Doctoral Dissertation

Essays on Trend Estimation by Penalized Least Squares

March,2021

Graduate School of Social Sciences,
Hiroshima University
Ruixue Du

Doctoral Dissertation

Essays on Trend Estimation by Penalized Least Squares

Supervisor: Hiroshi Yamada, Professor

Kazuhiko Hayakawa, Professor

March, 2021

Graduate School of Social Sciences,

Hiroshima university

Department of Economics

Ruixue Du

Acknowledgements

Two years of master and three and a half years of Ph.D. studies at Hiroshima University is the most important and meaningful period of my life. I would not be able to finish this journey without great assistance and supports from many people during these years. Therefore, please allow me to express my gratitude to those advisors and contributors.

First, my deepest gratitude will be expressed to my academic supervisor Prof. Hiroshi Yamada. I am grateful for his patient guidance, and enormous knowledge that makes me grow as a researcher. He gives me full supports not only with my research but also with my life. Prof. Hiroshi Yamada is very concerned about the life of international students. When I applied for a scholarship or tuition fee reduction, I asked him to write a recommendation or sign, and he always helped me without hesitation.

Besides my supervisor, I would like to thank my thesis committee members, Prof. Kazuhiko Hayakawa and Prof. Takashi Senda, for their sharp comments and valuable suggestions throughout my studies. Without their help, my thesis would not have been completed so smoothly and completely.

Last but not the least, I thank my parents for their endless love, encouragement, and supports, they always stand behind me as my strong backing. I would like to express my sincere gratitude to my dear sister for taking care of our parents when I have been studying in Japan.

Contents

Chapter 1

Introduction

In econometric analysis, trend estimation/smoothing methods based on penalized least squares are popular. This thesis focuses on three such methods. They are (a) Whittaker–Henderson (WH) method of graduation, which include Hodrick and Prescott (1997) filter as a special case, (b) ℓ_1 (polynomial) trend filtering developed by Kim et al. (1999), and (c) cubic smoothing spline, which was developed by Schoenberg (1964), Reinsch (1967) and others. In this chapter, we briefly review some researches which are closely related to our studies and then present the outline of the thesis.

1.1 Introductory survey

1.1.1 Bohlmann (1899): A pioneering study

Over 120 years ago, Bohlmann (1899) proposed the following trend estimation method:

$$\min_{x_1,x_2,\ldots,x_n\in\mathbb{R}} f(x_1,x_2,\ldots,x_n) = \sum_{i=1}^{n}(y_i-x_i)^2 + \lambda\sum_{i=1}^{n-1}(\nabla x_i)^2, \tag{1.1}$$

where $y_1,y_2,\ldots,y_n$ denote a time series, λ is a positive parameter, and $\nabla x_i = x_{i+1}-x_i$ for $i=1,2,\ldots,n-1$. $\sum_{i=1}^{n}(y_i-x_i)^2$ and $\sum_{i=1}^{n-1}(\nabla x_i)^2$ in (1.1) represent fidelity to the data and smoothness, respectively, and λ controls the trade-off between them.

Denote the solution of (1.1) by $\widehat{x}_1,\widehat{x}_2,\ldots,\widehat{x}_n$. More precisely, $\widehat{x}_1,\widehat{x}_2,\ldots,\widehat{x}_n$ are such that

$$f(x_1,x_2,\ldots,x_n) \geq f(\widehat{x}_1,\widehat{x}_2,\ldots,\widehat{x}_n). \tag{1.2}$$

Notably, Bohlmann (1899) obtained an explicit representation of $\widehat{x}_1, \widehat{x}_2, \ldots, \widehat{x}_n$. He showed

$$\widehat{x}_i = y_i + \lambda(\nabla\widehat{x}_i - \nabla\widehat{x}_{i-1}), \quad i = 1, 2, \ldots, n, \tag{1.3}$$

and

$$\begin{cases} \nabla\widehat{x}_0 = 0, \\ \nabla\widehat{x}_1 = \frac{\sum_{t=1}^{n-1} \nabla y_t \sinh\{(n-t)\alpha\}}{\lambda \sinh(n\alpha)}, \\ \nabla\widehat{x}_2 = \frac{\sinh\{(n-2)\alpha\}\nabla y_1 \sinh(\alpha) + \sinh 2\alpha \sum_{t=2}^{n-1} \nabla y_t \sinh\{(n-t)\alpha\}}{\lambda \sinh(\alpha)\sinh(n\alpha)}, \\ \quad\vdots \\ \nabla\widehat{x}_{n-1} = \frac{\sinh(\alpha)\sum_{t=1}^{n-2} \nabla y_t \sinh(t\alpha) + \sinh\{(n-1)\alpha\}\sum_{t=n-1}^{n-1} \nabla y_t \sinh\{(n-t)\alpha\}}{\lambda \sinh(\alpha)\sinh(n\alpha)}, \\ \nabla\widehat{x}_n = 0, \end{cases} \tag{1.4}$$

where $\alpha > 0$ is defined by

$$\cosh\alpha = 1 + \frac{1}{2\lambda}.$$

Here, we note that $\sinh$ and $\cosh$ represent the hyperbolic sine function and hyperbolic cos function, respectively. They are given by

$$\sinh(x) = \frac{e^x - e^{-x}}{2} \quad \text{and} \quad \cosh(x) = \frac{e^x + e^{-x}}{2}.$$

Then, $\widehat{x}_1, \widehat{x}_2, \ldots, \widehat{x}_n$ are obtainable by substituting (1.4) into (1.3).

We remark that proofs of (1.3) and (1.4) are provided in the Appendix. See Sections 1.3.1–1.3.2. A numerical example for obtaining $\widehat{x}_1, \widehat{x}_2, \ldots, \widehat{x}_n$ from (1.3) and (1.4) is also given in the Appendix. See Section 1.3.3.

1.1.2 Whittaker–Henderson method of graduation

Nearly a quarter of a century later, Whittaker (1923), without knowing Bohlmann (1899), proposed a similar idea to (1.1). Whittaker (1923) proposed the following method:

$$\min_{x_1,x_2,\ldots,x_n\in\mathbb{R}} f(x_1,x_2,\ldots,x_n) = \sum_{i=1}^{n} h_i^2(y_i - x_i)^2 + \lambda\sum_{i=1}^{n-3}\left(\nabla^3 x_i\right)^2, \tag{1.5}$$

where h_i $(i = 1, 2, \ldots, n)$ is a constant, ∇ represents the forward difference operator, and $\nabla^3 x_i = \nabla^2 x_{i+1} - \nabla^2 x_i$ denotes the third-order difference. The first term of (1.5) measures the closeness of fit (fidelity), and the second term is a measure of smoothness. Let $\epsilon = \dfrac{h_i^2}{\lambda}$ for $i = 4, 5, \ldots, n-3$, and $\widehat{x}_1, \widehat{x}_2, \ldots, \widehat{x}_n$ denote the solution of (1.5). Whittaker (1923) showed the following sixth-order difference equation holds:

$$\epsilon\widehat{x}_i = \epsilon y_i + \nabla^6\widehat{x}_{i-3}. \tag{1.6}$$

The proof of (1.6) is shown in the Appendix. See Section 1.3.4. Whittaker's student Aitken (1925) obtained an exact solution of Whittaker's equation above and almost the same time Henderson (1924) proposed a simplified way to solve the difference equation.

More than fifty years later, Hodrick and Prescott (1981, 1997) used a method similar to (1.1) and (1.5). That is the following minimization problem:

$$\min_{x_1,x_2,\ldots,x_n\in\mathbb{R}} f(x_1,x_2,\ldots,x_n) = \sum_{i=1}^{n}(y_i - x_i)^2 + \lambda\sum_{i=3}^{n}\left(\Delta^2 x_i\right)^2, \tag{1.7}$$

where Δ denotes the backward difference operator, and $\Delta^2 x_i = \Delta x_i - \Delta x_{i-1} = x_i - 2x_{i-1} + x_{i-2}$ is called the second-order difference. The first part in (1.7) is used to measure the fitness of the estimation to the original data, and the second part measures the smoothness. λ is a positive smoothing parameter to control the balance of smoothness and fitness. Their paper had a great impact on macroeconometric time series analysis and, in econometrics, (1.7) is referred to as "Hodrick-Prescott (HP) filter." In econometrics, a large volume of literature has been published in the last decade focusing on HP filter. Examples include Phillip and Jin (2015), de Jong and Sakarya (2016), Cornea-Madeira (2017), Hamilton (2018), Pillips and Shi (2019), Sakarya and de Jong (2020), Yamada (2015, 2018ab, 2020ab), Yamada and Du (2019), Yamada and Jahra (2019). Danthine and Girardin (1989) stated that (1.7) can be represented in matrix notation as follows:

$$\min_{\boldsymbol{x}\in\mathbb{R}} f(\boldsymbol{x}) = (\boldsymbol{y}-\boldsymbol{x})^\top(\boldsymbol{y}-\boldsymbol{x}) + \lambda\boldsymbol{x}^\top\boldsymbol{D}_2^\top\boldsymbol{D}_2\boldsymbol{x}, \tag{1.8}$$

where $\boldsymbol{y} = [y_1, y_2, \ldots, y_n]^\top$, $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]^\top$, and $\boldsymbol{D}_2 \in \mathbb{R}^{(n-2)\times n}$ is a difference matrix such that $\boldsymbol{D}_2\boldsymbol{x} = [\Delta^2 x_3, \ldots, \Delta^2 x_n]^\top$. They showed the solution of (1.8), denoted by $\widehat{\boldsymbol{x}}$, can be expressed by

$$\widehat{\boldsymbol{x}} = (\boldsymbol{I}_n + \lambda\boldsymbol{D}_2^\top\boldsymbol{D}_2)^{-1}\boldsymbol{y}, \tag{1.9}$$

where $\boldsymbol{I}_n \in \mathbb{R}^{n\times n}$ be an identity matrix. We provide Matlab/GNU Octave and R functions for calculating $\widehat{\boldsymbol{x}}$ in the Appendix. See Section 1.3.5.

As a generalization of (1.1), (1.5), and (1.7), consider the following minimization problem:

$$\min_{x_1,x_2,\ldots,x_n\in\mathbb{R}} f(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n}(y_i - x_i)^2 + \lambda\sum_{i=1}^{n-p}(\nabla^p x_i)^2, \tag{1.10}$$

where $\lambda > 0$, $0 < p < n$, and ∇^p denotes the p-th forward difference operator such that $\nabla^p x_i = \nabla^{p-1}x_{i+1} - \nabla^{p-1}x_i$. (1.10) is referred to as 'Whittaker–Henderson method of graduation.' See, e.g., Weinert (2007). The problem of (1.10) can be written in matrix notation as follows:

$$\min_{\boldsymbol{x}\in\mathbb{R}} f(\boldsymbol{x}) = (\boldsymbol{y}-\boldsymbol{x})^\top(\boldsymbol{y}-\boldsymbol{x}) + \lambda\boldsymbol{x}^\top\boldsymbol{D}_p^\top\boldsymbol{D}_p\boldsymbol{x}, \tag{1.11}$$

where $\boldsymbol{D}_p$ is a $(n-p)\times n$ difference matrix such that $\boldsymbol{D}_p\boldsymbol{x} = [\nabla^p x_1, \nabla^p x_2, \ldots, \nabla^p x_n]^\top$ for an n-dimension column vector $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]^\top$. The minimization of (1.11) is the solution of the following equation:

$$(\boldsymbol{I}_n + \lambda\boldsymbol{D}_p^\top\boldsymbol{D}_p)\widehat{\boldsymbol{x}} = \boldsymbol{y}. \tag{1.12}$$

1.1.3 ℓ_1 trend filtering

Kim et al. (2009) proposed a new method of trend estimation, "ℓ_1 trend filtering." The filtering method looks like HP filter. It is obtainable by replacing the squared ℓ_2 norm, $\sum_{i=3}^{n}(\Delta^2 x_i)^2$, in (1.7) with the ℓ_1 norm, $\sum_{i=3}^{n}|\Delta^2 x_i|$. More precisely, it is defined as

$$\begin{aligned}\widetilde{\boldsymbol{x}} &= \arg\min_{x_1,x_2,\dots,x_n} \sum_{i=1}^{n}(y_i - x_i)^2 + \lambda \sum_{i=3}^{n}|\Delta^2 x_i| \\ &= \arg\min_{\boldsymbol{x}\in\mathbb{R}^n} \|\boldsymbol{y}-\boldsymbol{x}\|_2^2 + \lambda\|\boldsymbol{D}_2\boldsymbol{x}\|_1, \end{aligned} \tag{1.13}$$

where λ is a positive tuning parameter, $\boldsymbol{y} = [y_1, y_2, \dots, y_n]^\top$, $\boldsymbol{x} = [x_1, x_2, \dots, x_n]^\top$, and $\boldsymbol{D}_2 \in \mathbb{R}^{(n-2)\times n}$ is a difference matrix such that $\boldsymbol{D}_2\boldsymbol{x} = [\Delta^2 x_3, \dots, \Delta^2 x_n]^\top$. In addition, $\|\boldsymbol{D}_2\boldsymbol{x}\|_1$ denotes the ℓ_1 norm of $\boldsymbol{D}_2\boldsymbol{x}$.

As the objective function of (1.13) is a coercive and strictly convex function, it has a unique global minimizer. Denote it by $\widetilde{\boldsymbol{x}}$. Concerning λ in (1.13), Kim et al. (2009) showed that

$$\begin{cases} \widetilde{\boldsymbol{x}} \to \boldsymbol{y} & \text{as } \lambda \to 0, \\ \boldsymbol{D}_2\widetilde{\boldsymbol{x}} = \boldsymbol{0} & \text{if } \lambda \geq \lambda_{\max}, \end{cases} \tag{1.14}$$

where

$$\lambda_{\max} = 2\|(\boldsymbol{D}_2\boldsymbol{D}_2^\top)^{-1}\boldsymbol{D}_2\boldsymbol{y}\|_\infty. \tag{1.15}$$

Here, for a vector $\boldsymbol{\eta} = [\eta_1, \dots, \eta_n]^\top$, $\|\boldsymbol{\eta}\|_\infty = \max\{|\eta_1|, \dots, |\eta_n|\}$. We remark that $\widetilde{\boldsymbol{x}}$ such that $\boldsymbol{D}_2\widetilde{\boldsymbol{x}} = \boldsymbol{0}$ represents a linear trend. This is because, in the case, $\widetilde{\boldsymbol{x}}$ belongs to the space spanned by $\boldsymbol{\iota} = [1, \dots, 1]^\top \in \mathbb{R}^n$ and $\boldsymbol{\tau} = [1, \dots, n]^\top \in \mathbb{R}^n$.

ℓ_1 trend filtering is attractive because it enables us to estimate a continuous piecewise linear trend. For the reason, it has been becoming popular in econometrics and finance. Examples include Yamada and Jin (2013), Yamada and Yoon (2014, 2016ab), Winkelried (2016), Yamada (2017ab), Klein (2018), and Mitra and Rohit (2018).

1.1.4 Smoothing spline

Cubic smoothing spline, which was developed by Schoenberg (1964), Reinsch (1967), and others, is a typical scatterplot smoothing method. Green and Silverman (1994) is an appropriate reference for it. As WH method of graduation and ℓ_1 trend filtering, it is a smoothing method based on penalized least-squares.

Consider the scatter plot of ordered pairs (x_i, y_i) for $i = 1, \dots, n$, where $x_1 < \cdots < x_n$. Let $\widehat{f}(x)$ represent the cubic smoothing spline whose knots are $x_1, \dots, x_n$ fitted to the same plot. More precisely, $\widehat{f}(x)$ is a function such that

$$\widehat{f}(x) = \arg \min_{f \in \mathcal{W}} \sum_{i=1}^{n} \{y_i - f(x_i)\}^2 + \lambda \int_a^b \left\{f''(x)\right\}^2 dx, \tag{1.16}$$

where a and b are such that $a < x_1$ and $x_n < b$, $\mathcal{W}$ denotes a function space contains all functions whose second derivative is square integrable over the interval $[a, b]$, and λ is a positive smoothing/-tuning parameter, which controls the trade-off between goodness of fit and smoothness.

Let $\widehat{\boldsymbol{f}} = [\widehat{f}(x_1), \dots, \widehat{f}(x_n)]^\top$. Then, as shown in Green and Silverman (1994), $\widehat{f}(x)$ is a natural cubic spline whose knots are $x_1, \dots, x_n$ and it thus follows that

$$\widehat{\boldsymbol{f}} = \arg \min_{\boldsymbol{f} \in \mathbb{R}^n} \|\boldsymbol{y} - \boldsymbol{f}\|^2 + \lambda \boldsymbol{f}^\top \boldsymbol{C}^\top \boldsymbol{R}^{-1} \boldsymbol{C} \boldsymbol{f} \tag{1.17}$$

$$= \left(\boldsymbol{I}_n + \lambda \boldsymbol{C}^\top \boldsymbol{R}^{-1} \boldsymbol{C}\right)^{-1} \boldsymbol{y}, \tag{1.18}$$

where $\boldsymbol{y} = [y_1, \dots, y_n]^\top$, $\boldsymbol{I}_n$ denotes the $n \times n$ identity matrix, and $\boldsymbol{C}$ and $\boldsymbol{R}$ are explicitly presented in Chapter 2. In addition, again as shown in Green and Silverman (1994), $\widehat{f}(x)$ in (1.16) is uniquely determined by $\widehat{\boldsymbol{f}} \in \mathbb{R}^n$ and therefore estimating $\widehat{f}(x)$ is equivalent to estimating $\widehat{\boldsymbol{f}}$.

Cubic smoothing spline has attracted a large amount of research attention in the last 30 years such as Speed (1991) indicated that fitting cubic smoothing spline is the best linear unbiased predictor (BLUP), many researchers have been working on the application and data analysis using cubic smoothing splines, including an approach to regression estimation (Cleveland and Devlin, 1988); local linear forecasts using cubic smoothing spline (Hyndman et al., 2002); the analysis of longitudinal data using cubic smoothing splines (Verbyla et al., 2012); solving a Cauchy problem using cubic smoothing spline (Nafuka et al., 2021).

1.2 Outline of the thesis

This thesis is organized as follows.

Chapter 2 is based on a research paper on cubic smoothing spline. Fitting a cubic smoothing spline is a typical smoothing method. In this study, we reveal a principle of duality in the penalized least squares regressions relating to the method. This is the main contribution of this study. We also provide a number of results derived from them, some of which are illustrated by a real data example.

Chapter 3 is based on a research paper on Whittaker–Henderson (WH) method of graduation. In the study, we present a modified WH method of graduation. After giving a closed-form solution, we show that it is of practical use because it provides not only a smoothed series identical to that of the WH graduation, but also an extrapolation beyond the sample limit of current data. In addition, we introduce two other penalized least squares problems and show that they provide the same results as those of the modified WH graduation.

Chapter 4 is based on a research paper on ℓ_1 polynomial trend filtering, which include ℓ_1 trend filtering as a special case. It is also a filtering method described as an ℓ_1-norm penalized least-squares problem. It is promising because it enables the estimation of a piecewise polynomial trend in a univariate economic time series without prespecifying the number and location of knots. This paper shows some theoretical results on the filtering, one of which is that a small modification of the filtering provides not only identical trend estimates as the filtering but also extrapolations of the trend beyond both sample limits.

1.3 Appendix

1.3.1 Proof of (1.3)

Let $\boldsymbol{D}_1$ denote a first-order difference matrix such that

$$\boldsymbol{D}_1 = \begin{bmatrix} -1 & 1 & 0 & 0 & \cdots & 0 \\ 0 & -1 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & -1 & 1 & 0 \\ 0 & \cdots & 0 & 0 & -1 & 1 \end{bmatrix} \in \mathbb{R}^{(n-1)\times n}. \tag{1.19}$$

The minimization problem in (1.1) can be represented in matrix notation as follows:

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}) = \|\boldsymbol{y}-\boldsymbol{x}\|_2^2+\lambda\|\boldsymbol{D}_1\boldsymbol{x}\|_2^2. \tag{1.20}$$

By differentiating $f(\boldsymbol{x})$ in (1.20) with respect to $\boldsymbol{x}$, we obtain

$$\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} = -2(\boldsymbol{y}-\boldsymbol{x}) + 2\lambda \boldsymbol{D}_1^\top \boldsymbol{D}_1 \boldsymbol{x}. \tag{1.21}$$

Let $\widehat{\boldsymbol{x}} = [\widehat{x}_1, \widehat{x}_2, \ldots, \widehat{x}_n]^\top$ denote the solution of (1.1) as before. Then, the optimal condition for (1.20) can be expressed by

$$-2(\boldsymbol{y}-\widehat{\boldsymbol{x}}) + 2\lambda \boldsymbol{D}_1^\top \boldsymbol{D}_1 \widehat{\boldsymbol{x}} = \boldsymbol{0}, \tag{1.22}$$

and accordingly we have

$$\boldsymbol{y}-\widehat{\boldsymbol{x}} = \lambda \boldsymbol{D}_1^\top \boldsymbol{D}_1 \widehat{\boldsymbol{x}}, \tag{1.23}$$

where

$$
\boldsymbol{D}_1^\top \boldsymbol{D}_1 = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & \cdots & 0 \\ -1 & 2 & -1 & 0 & 0 & \cdots & 0 \\ 0 & -1 & 2 & -1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & -1 & 2 & -1 & 0 \\ 0 & \cdots & 0 & 0 & -1 & 2 & -1 \\ 0 & \cdots & 0 & 0 & 0 & -1 & 1 \end{bmatrix} \in \mathbb{R}^{n\times n}. \tag{1.24}
$$

From (1.23)–(1.24), we have

$$
\begin{cases} y_1 - \widehat{x}_1 = \lambda(\widehat{x}_1 - \widehat{x}_2) = -\lambda(\nabla\widehat{x}_1 - \nabla\widehat{x}_0), \\ y_2 - \widehat{x}_2 = \lambda(-\widehat{x}_1 + 2\widehat{x}_2 - \widehat{x}_3) = -\lambda(\nabla\widehat{x}_2 - \nabla\widehat{x}_1), \\ \qquad\qquad \vdots \\ y_{n-1} - \widehat{x}_{n-1} = \lambda(-\widehat{x}_{n-2} + 2\widehat{x}_{n-1} - \widehat{x}_n) = -\lambda(\nabla\widehat{x}_{n-1} - \nabla\widehat{x}_{n-2}), \\ y_n - \widehat{x}_n = \lambda(\widehat{x}_n - \widehat{x}_{n-1}) = -\lambda(\nabla\widehat{x}_n - \nabla\widehat{x}_{n-1}). \end{cases} \tag{1.25}
$$

which leads to (1.3).

1.3.2 Proof of (1.4)

The solution of WH(1) filter in (1.20) is given by

$$
\widehat{\boldsymbol{x}} = (\boldsymbol{I}_n + \lambda \boldsymbol{D}_1^\top \boldsymbol{D}_1)^{-1} \boldsymbol{y}, \tag{1.26}
$$

from which we have

$$
(\boldsymbol{I}_n + \lambda \boldsymbol{D}_1^\top \boldsymbol{D}_1)\widehat{\boldsymbol{x}} = \boldsymbol{y}. \tag{1.27}
$$

Premultiplying (1.27) by $\boldsymbol{D}_1$ yields

$$\boldsymbol{D}_1\widehat{\boldsymbol{x}} = (\boldsymbol{I}_{n-1} + \lambda \boldsymbol{D}_1\boldsymbol{D}_1^\top)^{-1}\boldsymbol{D}_1\boldsymbol{y}. \tag{1.28}$$

Here, $(\boldsymbol{I}_{n-1} + \lambda \boldsymbol{D}_1\boldsymbol{D}_1^\top)$ is a symmetric tridiagonal matrix as follows:

$$\begin{bmatrix} 1+2\lambda & -\lambda & 0 & 0 & 0 & \cdots & 0 \\ -\lambda & 1+2\lambda & -\lambda & 0 & 0 & \cdots & 0 \\ 0 & -\lambda & 1+2\lambda & -\lambda & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & -\lambda & 1+2\lambda & -\lambda & 0 \\ 0 & \cdots & 0 & 0 & -\lambda & 1+2\lambda & -\lambda \\ 0 & \cdots & 0 & 0 & 0 & -\lambda & 1+2\lambda \end{bmatrix} \in \mathbb{R}^{(n-1)\times(n-1)}. \tag{1.29}$$

From Dow (2003, pp. E202–E203), the (i,j) element of $(\boldsymbol{I}_{n-1} + \lambda \boldsymbol{D}_1\boldsymbol{D}_1^\top)^{-1}$ is explicitly expressed as follows:

$$\boldsymbol{P}_{i,j} = \frac{\sinh(i\alpha)\sinh\{(n-j)\alpha\}}{\lambda\sinh(\alpha)\sinh(n\alpha)}, \quad i \leq j, \tag{1.30}$$

and

$$\boldsymbol{Q}_{i,j} = \frac{\sinh(j\alpha)\sinh\{(n-i)\alpha\}}{\lambda\sinh(\alpha)\sinh(n\alpha)}, \quad i > j, \tag{1.31}$$

where

$$\cosh(\alpha) = 1 + \frac{1}{2\lambda}.$$

Let $\boldsymbol{e}_i = [0,\ldots,0,1,0,\ldots,0]^\top \in \mathbb{R}^{n-1}$. Then, given that ∇x_i denotes the i-th entry of $\boldsymbol{D}_1\widehat{\boldsymbol{x}}$, we have

$$\nabla\widehat{x}_i = \boldsymbol{e}_i^\top(\boldsymbol{D}_1\widehat{\boldsymbol{x}}), \quad i = 1,2,\ldots,n-1. \tag{1.32}$$

By combining (1.28) and (1.32), it follows that

$$
\begin{aligned}
\nabla\widehat{x}_i &= \boldsymbol{e}_i^\top \boldsymbol{D}_1 \widehat{\boldsymbol{x}} = \boldsymbol{e}_i^\top \left(\boldsymbol{I}_{n-1} + \lambda \boldsymbol{D}_1 \boldsymbol{D}_1^\top\right)^{-1} \boldsymbol{D}_1 \boldsymbol{y} = \boldsymbol{e}_i^\top \left(\boldsymbol{I}_{n-1} + \lambda \boldsymbol{D}_1 \boldsymbol{D}_1^\top\right)^{-1} \boldsymbol{D}_1 \boldsymbol{y} \\
&= \boldsymbol{e}_i^\top \begin{bmatrix} \boldsymbol{P}_{1,1} & \boldsymbol{P}_{1,2} & \cdots & \boldsymbol{P}_{1,n-1} \\ \boldsymbol{Q}_{2,1} & \boldsymbol{P}_{2,2} & \cdots & \boldsymbol{P}_{2,n-1} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{Q}_{n-1,1} & \boldsymbol{Q}_{n-1,2} & \cdots & \boldsymbol{P}_{n-1,n-1} \end{bmatrix} \begin{bmatrix} \nabla y_1 \\ \nabla y_2 \\ \vdots \\ \nabla y_{n-1} \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{Q}_{i,1} & \boldsymbol{Q}_{i,2} & \cdots & \cdots & \boldsymbol{Q}_{i,i-1} \end{bmatrix} \begin{bmatrix} \nabla y_1 \\ \nabla y_2 \\ \vdots \\ \nabla y_{i-1} \end{bmatrix} \\
&\qquad + \begin{bmatrix} \boldsymbol{P}_{i,i} & \boldsymbol{P}_{i,i+1} & \cdots & \cdots & \boldsymbol{P}_{i,n-1} \end{bmatrix} \begin{bmatrix} \nabla y_i \\ \nabla y_{i+1} \\ \vdots \\ \nabla y_{n-1} \end{bmatrix}. \qquad (1.33)
\end{aligned}
$$

Thus, given (1.30) and (1.31), we obtain

$$
\begin{aligned}
\nabla\widehat{x}_i &= \left[\frac{\sinh(\alpha)\sinh\{(n-i)\alpha\}}{\lambda\sinh(\alpha)\sinh(n\alpha)} \quad \cdots \quad \frac{\sinh\{(i-1)\alpha\}\sinh\{(n-i)\alpha\}}{\lambda\sinh(\alpha)\sinh(n\alpha)}\right] \begin{bmatrix} \nabla y_1 \\ \nabla y_2 \\ \vdots \\ \nabla y_{i-1} \end{bmatrix} \\
&\quad + \left[\frac{\sinh(i\alpha)\sinh\{(n-i)\alpha\}}{\lambda\sinh(\alpha)\sinh(n\alpha)} \quad \cdots \quad \frac{\sinh(i\alpha)\sinh(\alpha)}{\lambda\sinh(\alpha)\sinh(n\alpha)}\right] \begin{bmatrix} \nabla y_i \\ \nabla y_{i+1} \\ \vdots \\ \nabla y_{n-1} \end{bmatrix} \\
&= \frac{\sinh\{(n-i)\alpha\}}{\lambda\sinh(\alpha)\sinh(n\alpha)} \left\{\sinh(\alpha)\nabla y_1 + \sinh(2\alpha)\nabla y_2 + \cdots + \sinh\{(i-1)\alpha\}\nabla y_{i-1}\right\} \\
&\quad + \frac{\sinh(i\alpha)}{\lambda\sinh(\alpha)\sinh(n\alpha)} \left\{\sinh\{(n-i)\alpha\}\nabla y_i + \sinh\{(n-i-1)\alpha\}\nabla y_{i+1} + \cdots + \sinh(\alpha)\nabla y_{n-1}\right\}
\end{aligned}
$$

$$= \frac{\sinh\{(n-i)\alpha\}}{\lambda \sinh(\alpha)\sinh(n\alpha)} \left\{ \sum_{t=1}^{i-1} \sinh(t\alpha)\nabla y_t \right\} + \frac{\sinh(i\alpha)}{\lambda \sinh(\alpha)\sinh(n\alpha)} \left\{ \sum_{t=i}^{n-1} \sinh\{(n-t)\alpha\}\nabla y_t \right\}$$
$$= \frac{\sinh\{(n-i)\alpha\}\sum_{t=1}^{i-1}\nabla y_t \sinh(t\alpha) + \sinh(i\alpha)\sum_{t=i}^{n-1}\nabla y_t \sinh\{(n-t)\alpha\}}{\lambda \sinh(\alpha)\sinh(n\alpha)}, \tag{1.34}$$

which leads to (1.4).

1.3.3 A numerical example for obtaining $\widehat{x}_1, \widehat{x}_2, \dots, \widehat{x}_n$ from (1.3) and (1.4)

As a numerical example, consider the case of such that $\boldsymbol{y} = [1, 3, 2, 4, 9, 5]^\top$ and $\lambda = 1$.
Then, given that $\cosh(\alpha) = 1 + \frac{1}{2\lambda} = \frac{3}{2}$, $\sinh(\alpha) = \frac{\sqrt{5}}{2}$, $\cosh(2\alpha) = \frac{7}{2}$, $\cosh(3\alpha) = 9$, $\sinh(2\alpha) = \frac{3\sqrt{5}}{2}$, $\sinh(3\alpha) = 4\sqrt{5}$, $\sinh(4\alpha) = \frac{21\sqrt{5}}{2}$, $\sinh(5\alpha) = \frac{55\sqrt{5}}{2}$, and $\sinh(6\alpha) = 72\sqrt{5}$, we have

$$\begin{aligned}
\nabla\widehat{x}_1 &= \frac{\sinh(\alpha)\sum_{t=1}^{5}\nabla y_t \sinh\{(6-t)\alpha\}}{\sinh(\alpha)\sinh(6\alpha)} \\
&= \frac{\nabla y_1 \sinh(5\alpha) + \nabla y_2 \sinh(4\alpha) + \nabla y_3 \sinh(3\alpha) + \nabla y_4 \sinh(2\alpha) + \nabla y_5 \sinh(\alpha)}{\sinh(6\alpha)} \\
&= \frac{2\sinh(5\alpha) - \sinh(4\alpha) + 2\sinh(3\alpha) + 5\sinh(2\alpha) - 4\sinh(\alpha)}{\sinh(6\alpha)} \\
&= \frac{\frac{55\sqrt{5}}{4} - 21\frac{\sqrt{5}}{2} + 8\frac{\sqrt{5}}{2} + 15\frac{\sqrt{5}}{2} - 4\frac{\sqrt{5}}{2}}{72\sqrt{5}} \\
&= \frac{29}{36} = 0.8056,
\end{aligned} \tag{1.35}$$

$$\begin{aligned}
\nabla\widehat{x}_2 &= \frac{\sinh(4\alpha)\sum_{t=1}^{1}\nabla y_t \sinh(t\alpha) + \sinh(2\alpha)\sum_{t=2}^{5}\nabla y_t \sinh\{(6-t)\alpha\}}{\sinh(\alpha)\sinh(6\alpha)} \\
&= \frac{2\sinh(4\alpha)\sinh(\alpha) + \sinh(2\alpha)\{-\sinh(4\alpha) + 2\sinh(3\alpha) + 5\sinh(2\alpha) - 4\sinh(\alpha)\}}{\sinh(\alpha)\sinh(6\alpha)} \\
&= \frac{2 \times \frac{21\sqrt{5}}{2} \times \frac{\sqrt{5}}{2} + \frac{3\sqrt{5}}{2}(-\frac{21\sqrt{5}}{2} + 8\sqrt{5} + 5 \times \frac{15\sqrt{5}}{2} - 2\sqrt{5})}{\frac{\sqrt{5}}{2} \times 72\sqrt{5}} \\
&= \frac{5}{12} = 0.4167,
\end{aligned} \tag{1.36}$$

$$\nabla\widehat{x}_3 = \frac{\sinh(3\alpha)\sum_{t=1}^{2}\nabla y_t \sinh(t\alpha) + \sinh(3\alpha)\sum_{t=3}^{5}\nabla y_t \sinh\{(6-t)\alpha\}}{\sinh(\alpha)\sinh(6\alpha)}$$

$$
\begin{aligned}
&= \frac{\sinh(3\alpha)\{2\sinh(\alpha) - \sinh(2\alpha)\} + \sinh(3\alpha)\{2\sinh(3\alpha) + 5\sinh(2\alpha) - 4\sinh(\alpha)\}}{\sinh(\alpha)\sinh(6\alpha)} \\
&= \frac{4\sqrt{5} \times (\sqrt{5} - \frac{3\sqrt{5}}{2}) + 4\sqrt{5} \times (8\sqrt{5} + \frac{15\sqrt{5}}{2} - 2\sqrt{5})}{\frac{\sqrt{5}}{2} \times 72\sqrt{5}} \\
&= \frac{13}{9} = 1.4444,
\end{aligned} \tag{1.37}
$$

$$
\begin{aligned}
\nabla\widehat{x}_4 &= \frac{\sinh(2\alpha)\sum_{t=1}^{3}\nabla y_t \sinh(t\alpha) + \sinh(4\alpha)\sum_{t=4}^{5}\nabla y_t \sinh\{(6-t)\alpha\}}{\sinh(\alpha)\sinh(6\alpha)} \\
&= \frac{\sinh(2\alpha)\{2\sinh(\alpha) - \sinh(2\alpha) + 2\sinh(3\alpha)\} + \sinh(4\alpha)\{2\sinh(5\alpha) - 4\sinh(\alpha)\}}{\sinh(\alpha)\sinh(6\alpha)} \\
&= \frac{\frac{3\sqrt{5}}{2} \times (\sqrt{5} - \frac{3\sqrt{5}}{2} + 8\sqrt{5}) + \frac{21\sqrt{5}}{2} \times (\frac{15\sqrt{5}}{2} - 2\sqrt{5})}{\frac{\sqrt{5}}{2} \times 72\sqrt{5}} \\
&= \frac{23}{12} = 1.9167,
\end{aligned} \tag{1.38}
$$

$$
\begin{aligned}
\nabla\widehat{x}_5 &= \frac{\sinh(\alpha)\sum_{t=1}^{4}\nabla y_t \sinh(t\alpha) + \sinh(5\alpha)\sum_{t=5}^{5}\nabla y_t \sinh\{(6-t)\alpha\}}{\sinh(\alpha)\sinh(6\alpha)} \\
&= \frac{\sinh(\alpha)\{2\sinh(\alpha) - \sinh(2\alpha) + 2\sinh(3\alpha) + 5\sinh(4\alpha)\} - 4\sinh(5\alpha)\sinh(\alpha)}{\sinh(\alpha)\sinh(6\alpha)} \\
&= \frac{\frac{\sqrt{5}}{2} \times (\sqrt{5} - \frac{3\sqrt{5}}{2} + 8\sqrt{5} + \frac{105\sqrt{5}}{2}) - 4 \times \frac{55\sqrt{5}}{2} \times \frac{\sqrt{5}}{2}}{\frac{\sqrt{5}}{2} \times 72\sqrt{5}} \\
&= -\frac{25}{36} = -0.6944.
\end{aligned} \tag{1.39}
$$

From (1.35)–(1.39), we obtain:

$$
\begin{aligned}
\boldsymbol{D}_1\widehat{\boldsymbol{x}} &= [\nabla x_1 - \nabla x_0, \nabla x_2 - \nabla x_1, \ldots, \nabla x_6 - \nabla x_5]^\top \\
&= [0.8056, -0.3889, 1.0278, 0.4722, -2.6111, 0.6944]^\top.
\end{aligned} \tag{1.40}
$$

Then, by using $\widehat{x}_i - y_i = \lambda(\nabla\widehat{x}_i - \nabla\widehat{x}_{i-1})$ for $i = 1, 2, \dots, n$, we obtain

$$\begin{aligned}\widehat{\boldsymbol{x}} &= [1 + 0.8056, 3 - 0.3889, 2 + 1.0278, 4 + 0.4722, 9 - 2.6111, 5 + 0.6944]^\top \\ &= [1.8056, 2.6111, 3.0278, 4.4722, 6.3889, 5.6944]^\top.\end{aligned} \tag{1.41}$$

We may confirm that these results are correct by using (1.28) and (1.26) as follows:

```
y=[1 3 2 4 9 5]';
lambda=1;
n=length(y);
D1=diff(eye(n));
D1xhat=inv(eye(n-1)+lambda*D1*D1')*D1*y

D1xhat =
    0.8056
    0.4167
    1.4444
    1.9167
   -0.6944

xhat=inv(eye(n)+lambda*D1'*D1)*y

xhat =
    1.8056
    2.6111
    3.0278
    4.4722
    6.3889
    5.6944
```

Here, we give another approach to obtain $\widehat{\boldsymbol{x}}$ from (1.3) and (1.4). Let $\boldsymbol{e}_1 = [1, 0, \dots, 0]^\top \in \mathbb{R}^n$ and $\boldsymbol{E} \in \mathbb{R}^{n\times n}$ be a matrix as follows:

$$\boldsymbol{E} = \begin{bmatrix} \boldsymbol{e}_1^\top \\ \boldsymbol{D}_1 \end{bmatrix}. \tag{1.42}$$

Then, it follows that

$$\boldsymbol{E}\widehat{\boldsymbol{x}} = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ -1 & 1 & 0 & 0 & \cdots & 0 \\ 0 & -1 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & -1 & 1 & 0 \\ 0 & \cdots & 0 & 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} \widehat{x}_1 \\ \widehat{x}_2 \\ \widehat{x}_3 \\ \vdots \\ \widehat{x}_{n-1} \\ \widehat{x}_n \end{bmatrix} = \begin{bmatrix} \widehat{x}_1 \\ \nabla\widehat{x}_1 \\ \nabla\widehat{x}_2 \\ \vdots \\ \nabla\widehat{x}_{n-2} \\ \nabla\widehat{x}_{n-1} \end{bmatrix}. \tag{1.43}$$

As $\boldsymbol{E}^{-1}$ is a lower triangular matrix of ones, $\widehat{\boldsymbol{x}}$ can be obtained by

$$\begin{bmatrix} \widehat{x}_1 \\ \widehat{x}_2 \\ \widehat{x}_3 \\ \vdots \\ \widehat{x}_{n-1} \\ \widehat{x}_n \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & \cdots & \cdots & 0 \\ 1 & 1 & 0 & \cdots & \cdots & 0 \\ 1 & 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\ 1 & 1 & 1 & \cdots & 1 & 0 \\ 1 & 1 & 1 & \cdots & 1 & 1 \end{bmatrix} \begin{bmatrix} \widehat{x}_1 \\ \nabla\widehat{x}_1 \\ \nabla\widehat{x}_2 \\ \vdots \\ \nabla\widehat{x}_{n-2} \\ \nabla\widehat{x}_{n-1} \end{bmatrix}. \tag{1.44}$$

1.3.4 Proof of (1.6)

Let $\boldsymbol{y} = [y_1, y_2, \ldots, y_n]^\top$, $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]^\top$, $\boldsymbol{H}$ and $\boldsymbol{D}_3$ be matrices such that

$$\boldsymbol{H} = \begin{bmatrix} h_1^2 & 0 & 0 & \cdots & 0 & 0 \\ 0 & h_2^2 & 0 & \cdots & 0 & 0 \\ 0 & 0 & h_3^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & h_{n-1}^2 & 0 \\ 0 & 0 & \cdots & 0 & 0 & h_n^2 \end{bmatrix} \in \mathbb{R}^{n\times n}, \tag{1.45}$$

$$\boldsymbol{D}_3 = \begin{bmatrix} -1 & 3 & -3 & 1 & 0 & 0 & \cdots & 0 \\ 0 & -1 & 3 & -3 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & -1 & 3 & -3 & 1 & 0 \\ 0 & \cdots & 0 & 0 & -1 & 3 & -3 & 1 \end{bmatrix} \in \mathbb{R}^{(n-3)\times n}. \tag{1.46}$$

The problem in (1.5) can be represented in matrix notation by

$$\min_{\boldsymbol{x}\in R^n} f(\boldsymbol{x}) = (\boldsymbol{y}-\boldsymbol{x})^\top \boldsymbol{H}(\boldsymbol{y}-\boldsymbol{x}) + \lambda(\boldsymbol{D}_3\boldsymbol{x})^\top(\boldsymbol{D}_3\boldsymbol{x}). \tag{1.47}$$

By differentiating $f(\boldsymbol{x})$ in (1.47) with respect to $\boldsymbol{x}$, we obtain

$$\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} = -2\boldsymbol{H}(\boldsymbol{y}-\boldsymbol{x}) + 2\lambda \boldsymbol{D}_3^\top \boldsymbol{D}_3 \boldsymbol{x}. \tag{1.48}$$

Let $\widehat{\boldsymbol{x}} = [\widehat{x}_1, \widehat{x}_2, \ldots, \widehat{x}_n]^\top$ denote the solution of (1.47), then the optimality condition for (1.47) can be expressed by

$$-\boldsymbol{H}(\boldsymbol{y}-\widehat{\boldsymbol{x}}) + \lambda \boldsymbol{D}_3^\top \boldsymbol{D}_3 \widehat{\boldsymbol{x}} = \boldsymbol{0}, \tag{1.49}$$

from which we have

$$\boldsymbol{H}\boldsymbol{y} = \boldsymbol{H}\widehat{\boldsymbol{x}} + \lambda \boldsymbol{D}_3^\top \boldsymbol{D}_3 \widehat{\boldsymbol{x}}. \tag{1.50}$$

Given $\boldsymbol{D}_3\widehat{\boldsymbol{x}} = [\nabla^3\widehat{x}_1, \nabla^3\widehat{x}_2, \ldots, \nabla^3\widehat{x}_{n-3}]^\top$, a set of equations can be derived:

$$\begin{cases} h_1^2 y_1 = h_1^2\widehat{x}_1 - \lambda\nabla^3\widehat{x}_1, \\ h_2^2 y_2 = h_2^2\widehat{x}_2 + 3\lambda\nabla^3\widehat{x}_1 - \lambda\nabla^3\widehat{x}_2, \\ h_3^2 y_3 = h_3^2\widehat{x}_3 - 3\lambda\nabla^3\widehat{x}_1 + 3\lambda\nabla^3\widehat{x}_2 - \lambda\nabla^3\widehat{x}_3, \\ h_4^2 y_4 = h_4^2\widehat{x}_4 + \lambda\nabla^3\widehat{x}_1 - 3\lambda\nabla^3\widehat{x}_2 + 3\lambda\nabla^3\widehat{x}_3 - \lambda\nabla^3\widehat{x}_4, \\ h_5^2 y_5 = h_5^2\widehat{x}_5 + \lambda\nabla^3\widehat{x}_2 - 3\lambda\nabla^3\widehat{x}_3 + 3\lambda\nabla^3\widehat{x}_4 - \lambda\nabla^3\widehat{x}_5, \\ h_6^2 y_6 = h_6^2\widehat{x}_6 + \lambda\nabla^3\widehat{x}_3 - 3\lambda\nabla^3\widehat{x}_4 + 3\lambda\nabla^3\widehat{x}_5 - \lambda\nabla^3\widehat{x}_6, \\ \qquad\quad\vdots \\ h_{n-3}^2 y_{n-3} = h_{n-3}^2\widehat{x}_{n-3} + \lambda\nabla^3\widehat{x}_{n-6} - 3\lambda\nabla^3\widehat{x}_{n-5} + 3\lambda\nabla^3\widehat{x}_{n-4} - \lambda\nabla^3\widehat{x}_{n-3}. \end{cases}$$

From the equations above, for $i = 4, 5, \ldots, n-3$, it follows that

$$h_i^2 y_i = h_i^2\widehat{x}_i + \lambda\nabla^3\widehat{x}_{i-3} - 3\lambda\nabla^3\widehat{x}_{i-2} + 3\lambda\nabla^3\widehat{x}_{i-1} - \lambda\nabla^3\widehat{x}_i. \tag{1.51}$$

Let $\epsilon = \dfrac{h_i^2}{\lambda}$ for $i = 4, 5, \ldots, n-3$, then (1.51) can be rewritten as

$$\epsilon y_i - \epsilon\widehat{x}_i = \nabla^3\widehat{x}_{i-3} - 3\nabla^3\widehat{x}_{i-2} + 3\nabla^3\widehat{x}_{i-1} - \nabla^3\widehat{x}_i. \tag{1.52}$$

Let $\widehat{x}_{i+1} = F\widehat{x}_i$, for the right-hand side of (1.52),

$$\begin{aligned} \nabla^3\widehat{x}_{i-3} - 3\nabla^3\widehat{x}_{i-2} + 3\nabla^3\widehat{x}_{i-1} - \nabla^3\widehat{x}_i &= \nabla^3(\widehat{x}_{i-3} - 3\widehat{x}_{i-2} + 3\widehat{x}_{i-1} - \widehat{x}_i) \\ &= \nabla^3(\widehat{x}_{i-3} - 3F\widehat{x}_{i-3} + 3F^2\widehat{x}_{i-3} - F^3\widehat{x}_{i-3}) \\ &= -\nabla^3(F-1)^3\widehat{x}_{i-3} \\ &= -\nabla^6\widehat{x}_{i-3}. \end{aligned} \tag{1.53}$$

From (1.52) and (1.53), the following sixth-order difference equation can be derived

$$\epsilon\widehat{x}_i = \epsilon y_i + \nabla^6\widehat{x}_{i-3}. \tag{1.54}$$

1.3.5 Matlab/R functions for calculating $\widehat{\boldsymbol{x}}$ in (1.9)

We give Matlab/R functions for calculating $\widehat{\boldsymbol{x}}$ in (1.9).

Matlab function:

```
function xhat=calcxhat(n,y,lambda)
   % n: sample size
   % lambda: smoothing parameter
   n = length(y);
   I = eye(n);
   D2 = diff(I,2);
   xhat = inv(I+lambda*D2'*D2)*y;
end
```

R function:

```
calcuxhat -> function(x){
   # x: n*1 vector
   n <- length(x);
   I <- diag(n);
   D2 <- diff(I,diff=2);
   xhat <- solve((I+lambda%*%t(D2)%*%D2))%*%y;
}
```

1.3.6 Matlab/R functions for calculating $\lambda_{\max}$ in (1.15)

We give Matlab/R functions for calculating $\lambda_{\max}$ in (1.15).

Matlab function:

```
function lambdamax=l1tf_lambdamax(y)
   % y: n*1 vector
   n=length(y);
   D2=diff(eye(n),2);
   lambdamax=norm((D2*D2')\(D2*y),inf);
   disp(sprintf('lambda_max : %e', lambdamax));
end
```

R function:

```
calculambdamax -> function(y){
   # y: n*1 vector
   n <- length(y);
   D2 <- diff(eye(n),diff=2);
   M <- solve(D2%*%t(D2))%*%(D2%*%y);
   lambdamax <- norm(M,p=inf);
}
```

1.4 References

1. Akiten, A. C., 1925, On the theory of graduation, *Proceedings of the Royal Society of Edinburgh*, 46, 36–45.

2. Bohlmann, G., 1899, Ein Ausgleichungsproblem, *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse*, 260–271.

3. Cleveland, W. S. and S. J. Devlin, 1988, Locally weighted regression: an approach to regression analysis by local fitting, *Journal of the American Statistical Association*, 83, 596–610.

4. Cornea-Madeira, A., 2017, The explicit formula for the Hodrick–Prescott filter in finite sample, *Review of Economics and Statistics*, 99, 2, 314–318.

5. Danthine, J.-P. and M. Girardin, 1989, Business cycles in Switzerland: A comparative study, *European Ecnomic Review*, 33, 1, 31–50.

6. De Jong, R. M. and N. Sakarya, 2016, The econometrics of the Hodrick–Prescott filter, *Review of Economics and Statistics*, 98, 2, 310–317.

7. Dow, M., 2003, Explicit inverses of Toeplitz and associated matrices, *ANZIM Jouranl*, 44, 185–215.

8. Green, P. J. and B. W. Silverman, 1994, *Nonparametric Regression and Generralized Linear Models: A Roughness Penalty Approach*, Chapman and Hall/CRC, Boca Raton.

9. Hamilton, J. D., 2018, Why you should never use the Hodrick-Prescott filter, *Review of Economics and Statistics*, 100, 831–843.

10. Henderson, R., 1924, A new method of graduation, *Transactions of the Actuarial Society of America*, 25, 29–40.

11. Henderson, R., 1925, Further remarks on graduation, *Transactions of the Actuarial Society of America*, 26, 52–57.

12. Hodrick, R. J. and E. C. Prescott, 1981, Post-war U.S. business cycles: An empirical investigation, Discussion Paper, No. 451, Center for Mathematical Studies in Economics and Management Science, Northwestern University.

13. Hodrick, R. J. and E. C. Prescott, 1997, Postwar U.S. business cycles: An empirical investigation, *Journal of Money, Credit and Banking*, 29, 1, 1–16.

14. Hyndman, R. J., A. B. Koehler, R. D. Snyder, and S. Grose, 2002, A state space framework for automatic forecasting using exponential smoothing methods, *International Journal of Forecasting*, 18, 3, 439–454.

15. Kim, S., K. Koh, S. Boyd, and D. Gorinevsky, 2009, ℓ_1 trend filtering, *SIAM Review*, 51, 2, 339–360.

16. Klein, T., 2018, Trend and contagion in WTI and brent crude oil spot and futures markets–The role of OPEC in the last decade, *Energy Economics*, 75, 636–646.

17. Mitra, S. K. and A. Rohit, 2020, Momentum trading with the ℓ_1-filter: Are the markets efficient?, *International Review of Finance*, 20, 4, 827–856.

18. Nanfuka, M., F. Berntsson, and G. Kakuba, 2021, Solving a cauchy problem for the heat equation using cubic smoothing splines, *Applicable Analisis*, `DOI:10.1080/00036811.2021.1876224`.

19. Phillips P. C. B. and S. Jin, 2015, Business cycles, trend elimination, and the HP filter, Cowles Discussion Paper, 2005, Yale University.

20. Phillips P. C. B. and Z. Shi, 2019, Boosting the Hodrick–Prescott filter, Cowles Foundation Discussion Paper, 2192, Yale University.

21. Reinsch, C., 1967, Smoothing by spline functions, *Numerische Maththematic*, 10, 177–183.

22. Sakarya, N. and R. M. De Jong, 2020, A property of the Hodrick–Prescott filter and its application, *Econometric Theory*, 36, 5, 840–870.

23. Schoenberg, I. J., 1964, Spline functions and the problem of graduation, *Proceedings of the National Academy of Sciences*, 52, 947–950.

24. Schuette, D. R., 1978, A linear programming approach to graduation, *Transactions of Society of Actuaries*, 30, 407–431.

25. Silverman, B. W., 1985, Some aspects of the spline smoothing approach to non-parametric regression curve fitting, *Journal of the Royal Statistical Society*, B, 47, 5–50.

26. Speed, T., 1991, Comment on paper by Robinson, *Statistical Science*, 6, 42–44.

27. Tibshirani, R. J., 2014, Adaptive piecewise polynomial estimation via trend filtering, *Annals of Statistics*, 42, 1, 285–323.

28. Verbyla, A. P., B. R. Cullis, M. G. Kenared, and S. J. Welham, The analysis of designed experiments and longitudinal data by using smooting splines, *Applied Statics*, 48, 3, 269–311.

29. Weinert, H. L., 2007, Efficient computation for Whittaker–Henderson smoothing, *Computational Statistics and Data Analysis*, 52, 2, 959–974.

30. Whittaker, E. T., 1923, On a new method of graduation, *Proceedings of the Edinburgh Mathematical Society*, 41, 63–75.

31. Winkelried, D., 2016, Piecewise linear trends and cycles in primary commodity prices, *Journal of International Money and Finance*, 64, 196–213.

32. Yamada, H., 2015, Ridge regression representions of the generalized Hodrick–Prescott filter, *Jouranl of the Japan statistical society*, 45, 2, 121–128.

33. Yamada, H., 2017a, Estimating the trend in US real GDP using the ℓ_1 trend filtering, *Applied Economics Letters*, 24, 10, 713–716.

34. Yamada, H., 2017b, A trend filtering method closely related to ℓ_1 trend filtering, *Empirical Economics*, 55,4, 1413–1423.

35. Yamada, H., 2018a, Why does the trend extracted by the Hodrick–Prescott filtering seem to be more plausible than the linear trend?, *Applied Economics Letters*, 25, 2, 102–105.

36. Yamada, H., 2018b, Several least squares problems related to the Hodrick–Prescott filtering, *Communications in Statistics – Theory and Methods*, 47, 5, 1022–1027.

37. Yamada, H., 2020a, A smoothing method that looks like the Hodick–Prescott filter, *Econometric Theory*, 36, 5, 1-21.

38. Yamada, H., 2020b, A pioneering study on discrete cosine transform, *Communications in Statistics – Theory and Methods*,
`DOI:10.1080/03610926.2020.1838547`.

39. Yamada, H. and R. Du, 2018, A modification of the Whittaker–Henderson method of graduation, *Communications in Statistics – Theory and Methods*, 48, 15, 3795–3800.

40. Yamada, H. and F. T. Jahra, 2019, Explicit formulars for the smoother weights of the Whittaker–Henderson graduation of order 1, *Communications in Statistics – Theory and Methods*, 48, 12, 3153–3161.

41. Yamada, H. and L. Jin, 2013, Japan's output gap estimation and ℓ_1 trend filtering, *Empirical Economics*, 45, 1, 81–88.

42. Yamada, H. and G. Yoon, 2014, When Grilli and Yang meet Prebisch and Singer: Piecewise linear trends in primary commodity prices, *Journal of International Money and Finance*, 42, 193–207.

43. Yamada, H. and G. Yoon, 2016a, Measuring the US NAIRU as a step function, *Empirical Economics*, 51, 4, 1679–1688.

44. Yamada, H. and G. Yoon, 2016b, Selecting the tuning parameter of the ℓ_1 trend filter, *Studies in Nonlinear Dynamics and Econometrics*, 20, 1, 97–105.

Chapter 2

Principle of Duality in Cubic Smoothing Spline

This chapter is based on a published article: Du and Yamada (2020).

2.1 Introduction

Fitting a cubic smoothing spline, which was developed by Schoenberg (1964), Reinsch (1967) and others, is a typical smoothing method. The cubic smoothing spline fitted to a scatter plot of ordered pairs (x_i, y_i) for $i = 1, \ldots, n$ is a function such that

$$\widehat{f}(x) = \arg\min_{f \in \mathcal{W}} \sum_{i=1}^{n} \{y_i - f(x_i)\}^2 + \lambda \int_a^b \left\{f''(x)\right\}^2 dx, \tag{2.1}$$

where $x_1, \ldots, x_n$ are points satisfying $a < x_1 < \cdots < x_n < b$, $\mathcal{W}$ denotes a function space that contains all functions whose second derivative is square integrable over $[a, b]$, and λ is a positive smoothing/tuning parameter, which controls the trade-off between goodness of fit and smoothness.

Let $\widehat{\boldsymbol{f}} = [\widehat{f}(x_1), \ldots, \widehat{f}(x_n)]^\top$. Then, given $\widehat{f}(x)$ is a natural cubic spline whose knots are $x_1, \ldots, x_n$ (see, e.g., Green and Silverman, 1994; Wood, 2017), it follows that

$$\begin{aligned} \widehat{\boldsymbol{f}} &= \arg\min_{\boldsymbol{f} \in \mathbb{R}^n} \|\boldsymbol{y} - \boldsymbol{f}\|^2 + \lambda \boldsymbol{f}^\top \boldsymbol{C}^\top \boldsymbol{R}^{-1} \boldsymbol{C} \boldsymbol{f} && (2.2) \\ &= \left(\boldsymbol{I}_n + \lambda \boldsymbol{C}^\top \boldsymbol{R}^{-1} \boldsymbol{C}\right)^{-1} \boldsymbol{y}, && (2.3) \end{aligned}$$

where $\boldsymbol{y} = [y_1, \dots, y_n]^\top$, $\boldsymbol{I}_n$ denotes the $n \times n$ identity matrix, and $\boldsymbol{C}$ and $\boldsymbol{R}$ are explicitly presented later. Then, as shown in Green and Silverman (1994), $\widehat{f}(x)$ in (2.1) is uniquely determined by $\widehat{\boldsymbol{f}} \in \mathbb{R}^n$ in (2.3). Thus, estimating $\widehat{f}(x)$ is equivalent to estimating $\widehat{\boldsymbol{f}}$.

Let $\boldsymbol{\Pi} = [\boldsymbol{\iota}_n, \boldsymbol{x}] \in \mathbb{R}^{n \times 2}$, where $\boldsymbol{\iota}_n = [1, \dots, 1]^\top \in \mathbb{R}^n$ and $\boldsymbol{x} = [x_1, \dots, x_n]^\top$. Note that since $x_1 < \cdots < x_n$, $\boldsymbol{\iota}_n$ and $\boldsymbol{x}$ are linearly independent and thus $\boldsymbol{\Pi}$ is of full column rank. Let

$$\widehat{\boldsymbol{\tau}} = \boldsymbol{\Pi}(\boldsymbol{\Pi}^\top \boldsymbol{\Pi})^{-1} \boldsymbol{\Pi}^\top \boldsymbol{y}. \tag{2.4}$$

Denote the difference between $\widehat{\boldsymbol{f}}$ and $\widehat{\boldsymbol{\tau}}$ (resp. $\boldsymbol{y}$ and $\widehat{\boldsymbol{f}}$) by $\widehat{\boldsymbol{c}}$ (resp. $\widehat{\boldsymbol{u}}$):

$$\widehat{\boldsymbol{c}} = \widehat{\boldsymbol{f}} - \widehat{\boldsymbol{\tau}}, \quad \widehat{\boldsymbol{u}} = \boldsymbol{y} - \widehat{\boldsymbol{f}}. \tag{2.5}$$

Accordingly, we have

$$\boldsymbol{y} = \widehat{\boldsymbol{\tau}} + \widehat{\boldsymbol{c}} + \widehat{\boldsymbol{u}}. \tag{2.6}$$

In this chapter, we present a comprehensive list of penalized least squares regressions relating to (2.6). One such example is the ridge regression (Hoerl and Kennard, 1970) that leads to $\widehat{\boldsymbol{c}}$. Then, we reveal a principle of duality in them. In addition, based on them, we provide a number of theoretical results, e.g., $\boldsymbol{\iota}_n^\top \widehat{\boldsymbol{c}} = 0$.

This chapter is organized as follows. Section 2.2 fixes some notations and gives key preliminary results used to derive the main results in the chapter. Section 2.3 provides a comprehensive list of penalized least squares regressions relating to (2.6), and reveals a principle of duality in them. Section 2.4 shows some results that are obtainable from the regressions shown in Section 2.3. Section 2.5 illustrates some results provided in Sections 2.3 and 2.4 by a real data example. Section 2.6 deals with the cases such that the other right-inverse matrices are used. Section 2.7 concludes the chapter.

2.2 Preliminaries

In this section, we give key preliminary results used to derive the main results of this chapter. Before stating them, we fix some notations.

2.2.1 Notations

Let $\widehat{f}_i$ (resp. $\widehat{\tau}_i$) denote the ith entry of $\widehat{\boldsymbol{f}}$ (resp. $\widehat{\boldsymbol{\tau}}$) for $i = 1, \ldots, n$, $\delta_i = x_{i+1} - x_i$, which is positive by definition, for $i = 1, \ldots, n-1$, $\boldsymbol{\Delta} = \mathsf{diag}(\delta_1, \ldots, \delta_{n-1}) \in \mathbb{R}^{(n-1)\times(n-1)}$, and for a full-*row*-rank matrix $\boldsymbol{M} \in \mathbb{R}^{m\times n}$, $\boldsymbol{M}^\top(\boldsymbol{M}\boldsymbol{M}^\top)^{-1} \in \mathbb{R}^{n\times m}$, which is a right-inverse matrix of $\boldsymbol{M}$, be denoted by $\boldsymbol{M}_{\mathrm{r}}^{-1}$. For a full-*column*-rank matrix $\boldsymbol{W} \in \mathbb{R}^{n\times p}$, let $\mathcal{S}(\boldsymbol{W})$ [resp. $\mathcal{S}^\perp(\boldsymbol{W})$] denote the column space of $\boldsymbol{W}$ [resp. the orthogonal complement of $\mathcal{S}(\boldsymbol{W})$] and $\boldsymbol{P}_W$ [resp. $\boldsymbol{Q}_W$] denote the orthogonal projection matrix to the space $\mathcal{S}(\boldsymbol{W})$ [resp. $\mathcal{S}^\perp(\boldsymbol{W})$]. Explicitly, they are $\boldsymbol{P}_W = \boldsymbol{W}(\boldsymbol{W}^\top\boldsymbol{W})^{-1}\boldsymbol{W}^\top$ and $\boldsymbol{Q}_W = \boldsymbol{I}_n - \boldsymbol{P}_W$. $\boldsymbol{D}_{(i)} \in \mathbb{R}^{(n-i)\times(n-i+1)}$ is a Toeplitz matrix whose first (resp. last) row is $[-1, 1, 0, \ldots, 0]$ (resp. $[0, \ldots, 0, -1, 1]$) and we define matrices $\boldsymbol{C} \in \mathbb{R}^{(n-2)\times n}$ and $\boldsymbol{R} \in \mathbb{R}^{(n-2)\times(n-2)}$ as follows:

$$\boldsymbol{C} = \left[\begin{array}{c:c:c:c:c:c}
\delta_1^{-1} & -\delta_1^{-1} - \delta_2^{-1} & \delta_2^{-1} & 0 & \cdots & 0 \\ \hdashline
0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \hdashline
\vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \hdashline
0 & \cdots & 0 & \delta_{n-2}^{-1} & -\delta_{n-2}^{-1} - \delta_{n-1}^{-1} & \delta_{n-1}^{-1}
\end{array}\right] \tag{2.7}$$

and

$$\boldsymbol{R} = \left[\begin{array}{c:c:c:c:c}
\frac{1}{3}(\delta_1 + \delta_2) & \frac{1}{6}\delta_2 & 0 & \cdots & 0 \\ \hdashline
\frac{1}{6}\delta_2 & \frac{1}{3}(\delta_2 + \delta_3) & \ddots & \ddots & \vdots \\ \hdashline
0 & \ddots & \ddots & \ddots & 0 \\ \hdashline
\vdots & \ddots & \ddots & \ddots & \frac{1}{6}\delta_{n-2} \\ \hdashline
0 & \cdots & 0 & \frac{1}{6}\delta_{n-2} & \frac{1}{3}(\delta_{n-2} + \delta_{n-1})
\end{array}\right]. \tag{2.8}$$

Finally, we denote the eigenvalues of $\boldsymbol{R}$ by $\omega_1, \ldots, \omega_{n-2}$ in descending order.

2.2.2 Key preliminary results

Lemma 2.1. *(i) $\boldsymbol{C}$ can be factorized as $\boldsymbol{C} = \boldsymbol{D}_{(2)}\boldsymbol{\Delta}^{-1}\boldsymbol{D}_{(1)}$. (ii) We have the following inequalities:*

$$\omega_{n-2} \geq \min\left\{\frac{1}{3}\delta_1 + \frac{1}{6}\delta_2, \frac{1}{6}(\delta_2+\delta_3), \dots, \frac{1}{6}(\delta_{n-3}+\delta_{n-2}), \frac{1}{6}\delta_{n-2} + \frac{1}{3}\delta_{n-1}\right\} > 0.$$

Proof of Lemma 2.1. (i) Let $\boldsymbol{w} = [w_1, \dots, w_n]^\top$ be an n-dimensional column vector. Then, by definition of $\boldsymbol{C}$, it follows that

$$\begin{aligned}\boldsymbol{C}\boldsymbol{w} &= \begin{bmatrix} -\frac{w_2-w_1}{\delta_1} + \frac{w_3-w_2}{\delta_2} \\ \vdots \\ -\frac{w_{n-1}-w_{n-2}}{\delta_{n-2}} + \frac{w_n-w_{n-1}}{\delta_{n-1}} \end{bmatrix} = \boldsymbol{D}_{(2)} \begin{bmatrix} \frac{w_2-w_1}{\delta_1} \\ \vdots \\ \frac{w_n-w_{n-1}}{\delta_{n-1}} \end{bmatrix} = \boldsymbol{D}_{(2)}\boldsymbol{\Delta}^{-1} \begin{bmatrix} -w_1+w_2 \\ \vdots \\ -w_{n-1}+w_n \end{bmatrix} \\ &= \boldsymbol{D}_{(2)}\boldsymbol{\Delta}^{-1}\boldsymbol{D}_{(1)}\boldsymbol{w} \in \mathbb{R}^{n-2},\end{aligned}$$

which leads to $\boldsymbol{C} = \boldsymbol{D}_{(2)}\boldsymbol{\Delta}^{-1}\boldsymbol{D}_{(1)}$. (ii) The first inequality follows by applying the Gershgorin circle theorem and the second inequality holds from $\delta_i > 0$ for $i = 1, \dots, n-1$. □

Remark 2.1. In the Appendix, we give some remarks on a special case such that $\boldsymbol{x} = [1, \dots, n]^\top$.

Lemma 2.2. *(i) $\mathcal{S}(\boldsymbol{C}^\top)$ equals $\mathcal{S}^\perp(\boldsymbol{\Pi})$ and (ii) $\mathcal{S}(\boldsymbol{C}_{\mathrm{r}}^{-1})$ equals $\mathcal{S}^\perp(\boldsymbol{\Pi})$.*

Proof of Lemma 2.2. (i) Given that $\delta_i > 0$ for $i = 1, \dots, n-1$, both $\boldsymbol{\Pi}$ and $\boldsymbol{C}^\top$ are of full column rank. In addition, $[\boldsymbol{\Pi}, \boldsymbol{C}^\top]$ is a square matrix. Thus, if $(\boldsymbol{C}^\top)^\top\boldsymbol{\Pi} = \boldsymbol{C}\boldsymbol{\Pi} = \boldsymbol{0}$, then it follows that $\mathcal{S}(\boldsymbol{C}^\top) = \mathcal{S}^\perp(\boldsymbol{\Pi})$. From $\boldsymbol{D}_{(1)}\boldsymbol{\iota}_n = \boldsymbol{0}$, we have $\boldsymbol{C}\boldsymbol{\iota}_n = \boldsymbol{D}_{(2)}\boldsymbol{\Delta}^{-1}\boldsymbol{D}_{(1)}\boldsymbol{\iota}_n = \boldsymbol{0}$. Likewise, from $\boldsymbol{\Delta}^{-1}\boldsymbol{D}_{(1)}\boldsymbol{x} = \boldsymbol{\Delta}^{-1}\boldsymbol{\Delta}\boldsymbol{\iota}_{n-1} = \boldsymbol{\iota}_{n-1}$ and $\boldsymbol{D}_{(2)}\boldsymbol{\iota}_{n-1} = \boldsymbol{0}$, we obtain $\boldsymbol{C}\boldsymbol{x} = \boldsymbol{D}_{(2)}\boldsymbol{\Delta}^{-1}\boldsymbol{D}_{(1)}\boldsymbol{x} = \boldsymbol{0}$. Accordingly, we have $\boldsymbol{C}\boldsymbol{\Pi} = \boldsymbol{0}$, which completes the proof. (ii) Recall that $\boldsymbol{C}_{\mathrm{r}}^{-1} = \boldsymbol{C}^\top(\boldsymbol{C}\boldsymbol{C}^\top)^{-1}$. It is clear that $\boldsymbol{C}_{\mathrm{r}}^{-1}$ is a full-column-rank matrix such that $[\boldsymbol{\Pi}, \boldsymbol{C}_{\mathrm{r}}^{-1}]$ is a square matrix. In addition, $(\boldsymbol{C}_{\mathrm{r}}^{-1})^\top\boldsymbol{\Pi} = (\boldsymbol{C}\boldsymbol{C}^\top)^{-1}\boldsymbol{C}\boldsymbol{\Pi} = \boldsymbol{0}$. Thus, it follows that $\mathcal{S}(\boldsymbol{C}_{\mathrm{r}}^{-1}) = \mathcal{S}^\perp(\boldsymbol{\Pi})$. □

Denote the spectral decomposition of $\boldsymbol{R}$ by $\boldsymbol{V}\boldsymbol{\Omega}\boldsymbol{V}^\top$ and let $\boldsymbol{R}^{-1/2} = \boldsymbol{V}\boldsymbol{\Omega}^{-1/2}\boldsymbol{V}^\top$, where $\boldsymbol{\Omega}^{-1/2} = \mathsf{diag}\left(1/\sqrt{\omega_1}, \dots, 1/\sqrt{\omega_{n-2}}\right)$. Then, $\boldsymbol{R}^{-1/2}$ is a positive definite matrix such that $\boldsymbol{R}^{-1/2}\boldsymbol{R}^{-1/2} = \boldsymbol{R}^{-1}$. Define

$$\boldsymbol{D} = \boldsymbol{R}^{-1/2}\boldsymbol{C}. \tag{2.9}$$

Then, given that $\boldsymbol{C}^\top$ is of full column rank and $\boldsymbol{R}^{-1/2}$ is nonsingular, $\boldsymbol{D} \in \mathbb{R}^{(n-2)\times n}$ is also of full row rank. In addition, we have

$$\boldsymbol{D}^\top \boldsymbol{D} = \boldsymbol{C}^\top \boldsymbol{R}^{-1} \boldsymbol{C}. \tag{2.10}$$

(We provide Matlab/GNU Octave and R functions for calculating $\boldsymbol{C}$, $\boldsymbol{R}$, and $\boldsymbol{D}$ in the Appendix.)

Lemma 2.3. *(i) $\mathcal{S}(\boldsymbol{D}^\top)$ equals $\mathcal{S}^\perp(\boldsymbol{\Pi})$ and (ii) $\mathcal{S}(\boldsymbol{D}_\mathrm{r}^{-1})$ equals $\mathcal{S}^\perp(\boldsymbol{\Pi})$.*

Proof of Lemma 2.3. Both (i) and (ii) may be proved similarly to Lemma 2.2(ii). For example, given $\boldsymbol{C\Pi} = \boldsymbol{0}$, we have $(\boldsymbol{D}^\top)^\top \boldsymbol{\Pi} = \boldsymbol{D\Pi} = \boldsymbol{R}^{-1/2}\boldsymbol{C\Pi} = \boldsymbol{0}$. □

Denote the eigenvalues of $\boldsymbol{C}^\top \boldsymbol{R}^{-1} \boldsymbol{C}$ by $g_1, \ldots, g_n$ in ascending order and the spectral decomposition of $\boldsymbol{C}^\top \boldsymbol{R}^{-1} \boldsymbol{C}$ by $\boldsymbol{UGU}^\top$, where $\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n]$ and $\boldsymbol{G} = \mathsf{diag}(g_1, \ldots, g_n)$. Let $\boldsymbol{T} = [\boldsymbol{u}_1, \boldsymbol{u}_2] \in \mathbb{R}^{n\times 2}$, $\boldsymbol{E}^\top = [\boldsymbol{u}_3, \ldots, \boldsymbol{u}_n] \in \mathbb{R}^{n\times(n-2)}$, and $\boldsymbol{S} = \mathsf{diag}(g_3, \ldots, g_n) \in \mathbb{R}^{(n-2)\times(n-2)}$.

Lemma 2.4. *(i) $\mathcal{S}(\boldsymbol{T})$ equals $\mathcal{S}(\boldsymbol{\Pi})$, (ii) $\mathcal{S}(\boldsymbol{E}^\top)$ equals $\mathcal{S}^\perp(\boldsymbol{\Pi})$, and (iii) $\mathcal{S}(\boldsymbol{E}_\mathrm{r}^{-1})$ equals $\mathcal{S}^\perp(\boldsymbol{\Pi})$.*

Proof of Lemma 2.4. (i) Since $\boldsymbol{C}^\top \boldsymbol{R}^{-1} \boldsymbol{C} \in \mathbb{R}^{n\times n}$ is a nonnegative definite matrix whose rank is $n-2$, we have $0 = g_1 = g_2 < g_3 < \cdots < g_n$. In addition, given $\boldsymbol{C\Pi} = \boldsymbol{0}$, it follows that $\boldsymbol{C}^\top \boldsymbol{R}^{-1} \boldsymbol{C\Pi} = 0 \cdot \boldsymbol{\Pi}$, which completes the proof. (ii) and (iii) They may be proved similarly to Lemma 2.2(ii). □

Given $g_1 = g_2 = 0$, we have

$$\boldsymbol{E}^\top \boldsymbol{SE} = \boldsymbol{C}^\top \boldsymbol{R}^{-1} \boldsymbol{C}. \tag{2.11}$$

Define

$$\boldsymbol{F} = \boldsymbol{S}^{1/2} \boldsymbol{E}, \tag{2.12}$$

where $\boldsymbol{S}^{1/2} = \mathsf{diag}(\sqrt{g_3}, \ldots, \sqrt{g_n}) \in \mathbb{R}^{(n-2)\times(n-2)}$. Then, we have

$$\boldsymbol{F}^\top \boldsymbol{F} = \boldsymbol{C}^\top \boldsymbol{R}^{-1} \boldsymbol{C}. \tag{2.13}$$

Lemma 2.5. *(i) $\mathcal{S}(\boldsymbol{F}^\top)$ equals $\mathcal{S}^\perp(\boldsymbol{\Pi})$ and (ii) $\mathcal{S}(\boldsymbol{F}_{\mathrm{r}}^{-1})$ equals $\mathcal{S}^\perp(\boldsymbol{\Pi})$.*

Proof of Lemma 2.5. Both (i) and (ii) may be proved similarly to Lemma 2.2(ii). For example, given $\boldsymbol{E}\boldsymbol{\Pi} = \boldsymbol{0}$, we have $(\boldsymbol{F}^\top)^\top\boldsymbol{\Pi} = \boldsymbol{F}\boldsymbol{\Pi} = \boldsymbol{S}^{1/2}\boldsymbol{E}\boldsymbol{\Pi} = \boldsymbol{0}$. □

Lemma 2.6. *There exists an orthogonal matrix $\boldsymbol{\Upsilon} \in \mathbb{R}^{(n-2)\times(n-2)}$ such that $\boldsymbol{F}^\top = \boldsymbol{D}^\top\boldsymbol{\Upsilon}$.*

Proof of Lemma 2.6. Recall that both $\boldsymbol{D}^\top \in \mathbb{R}^{n\times(n-2)}$ and $\boldsymbol{F}^\top \in \mathbb{R}^{n\times(n-2)}$ are of full column rank and $\mathcal{S}(\boldsymbol{D}^\top) = \mathcal{S}(\boldsymbol{F}^\top)$. Accordingly, these exists a nonsingular matrix $\boldsymbol{\Upsilon} \in \mathbb{R}^{(n-2)\times(n-2)}$ such that $\boldsymbol{F}^\top = \boldsymbol{D}^\top\boldsymbol{\Upsilon}$. Given that $\boldsymbol{D}^\top\boldsymbol{D} = \boldsymbol{F}^\top\boldsymbol{F}$, we have $\boldsymbol{D}^\top(\boldsymbol{I}_{n-2} - \boldsymbol{\Upsilon}\boldsymbol{\Upsilon}^\top)\boldsymbol{D} = \boldsymbol{0}$. Then, from $\boldsymbol{D}_{\mathrm{r}}^{-1\top}\boldsymbol{D}^\top(\boldsymbol{I}_{n-2} - \boldsymbol{\Upsilon}\boldsymbol{\Upsilon}^\top)\boldsymbol{D}\boldsymbol{D}_{\mathrm{r}}^{-1} = \boldsymbol{I}_{n-2} - \boldsymbol{\Upsilon}\boldsymbol{\Upsilon}^\top = \boldsymbol{0}$, we have $\boldsymbol{\Upsilon}^\top = \boldsymbol{\Upsilon}^{-1}$. □

Let (i) $\mathcal{A} = \boldsymbol{D}, \boldsymbol{F}$, (ii) $(\mathcal{B}, \mathcal{Q}) = (\boldsymbol{C}, \boldsymbol{R}), (\boldsymbol{E}, \boldsymbol{S}^{-1})$, (iii) $\mathcal{D} = \boldsymbol{C}, \boldsymbol{D}, \boldsymbol{E}, \boldsymbol{F}$, and (iv) $\mathcal{P} = \boldsymbol{\Pi}, \boldsymbol{T}$. From the results above, we immediately obtain the following results:

Proposition 2.1. *(i) $\boldsymbol{C}^\top\boldsymbol{R}^{-1}\boldsymbol{C} = \mathcal{A}^\top\mathcal{A} = \mathcal{B}^\top\mathcal{Q}^{-1}\mathcal{B}$, (ii) $\mathcal{D}\mathcal{P} = \mathcal{D}_{\mathrm{r}}^{-1\top}\mathcal{P} = \boldsymbol{0}$, (iii) both $[\mathcal{P}, \mathcal{D}^\top]$ and $[\mathcal{P}, \mathcal{D}_{\mathrm{r}}^{-1}]$ are nonsingular, and (iv) $\boldsymbol{P}_{\mathcal{D}^\top} = \boldsymbol{P}_{\mathcal{D}_{\mathrm{r}}^{-1}} = \boldsymbol{Q}_{\mathcal{P}}$.*

2.3 Several regressions relating to (2.6) and principle of duality in them

In this section, we provide a comprehensive list of penalized least squares regressions relating to (2.6), and reveal a principle of duality in them. The penalized regressions are, more precisely, those to compute $\widehat{\boldsymbol{c}}$, $\widehat{\boldsymbol{u}}$, $\widehat{\boldsymbol{\tau}}$, $\widehat{\boldsymbol{\tau}} + \widehat{\boldsymbol{c}}$, $\widehat{\boldsymbol{c}} + \widehat{\boldsymbol{u}}$, and $\widehat{\boldsymbol{\tau}} + \widehat{\boldsymbol{u}}$.

2.3.1 Penalized regressions to compute $\widehat{\boldsymbol{\tau}} + \widehat{\boldsymbol{c}}$

Concerning $\widehat{\boldsymbol{\tau}} + \widehat{\boldsymbol{c}}$, we have the following results:

Lemma 2.7. *It follows that*

$$\widehat{\boldsymbol{\tau}} + \widehat{\boldsymbol{c}} = \arg\min_{\boldsymbol{f}\in\mathbb{R}^n} \|\boldsymbol{y} - \boldsymbol{f}\|^2 + \lambda\|\mathcal{A}\boldsymbol{f}\|^2 = \left(\boldsymbol{I}_n + \lambda\mathcal{A}^\top\mathcal{A}\right)^{-1}\boldsymbol{y} \tag{2.14}$$

$$= \arg\min_{\boldsymbol{f}\in\mathbb{R}^n} \|\boldsymbol{y} - \boldsymbol{f}\|^2 + \lambda\boldsymbol{f}^\top\mathcal{B}^\top\mathcal{Q}^{-1}\mathcal{B}\boldsymbol{f} = \left(\boldsymbol{I}_n + \lambda\mathcal{B}^\top\mathcal{Q}^{-1}\mathcal{B}\right)^{-1}\boldsymbol{y}. \tag{2.15}$$

Proof of Lemma 2.7. From Proposition 2.1, we have $\boldsymbol{C}^\top \boldsymbol{R}^{-1}\boldsymbol{C} = \mathcal{A}^\top\mathcal{A} = \mathcal{B}^\top\mathcal{Q}^{-1}\mathcal{B}$. Then, (2.2)–(2.3) can be represented as follows:

$$
\begin{aligned}
\widehat{\boldsymbol{f}} &= \arg\min_{\boldsymbol{f}\in\mathbb{R}^n} \|\boldsymbol{y}-\boldsymbol{f}\|^2 + \lambda\|\mathcal{A}\boldsymbol{f}\|^2 = \left(\boldsymbol{I}_n + \lambda\mathcal{A}^\top\mathcal{A}\right)^{-1}\boldsymbol{y} \\
&= \arg\min_{\boldsymbol{f}\in\mathbb{R}^n} \|\boldsymbol{y}-\boldsymbol{f}\|^2 + \lambda\boldsymbol{f}^\top\mathcal{B}^\top\mathcal{Q}^{-1}\mathcal{B}\boldsymbol{f} = \left(\boldsymbol{I}_n + \lambda\mathcal{B}^\top\mathcal{Q}^{-1}\mathcal{B}\right)^{-1}\boldsymbol{y}.
\end{aligned}
$$

In addition, by definition of $\widehat{\boldsymbol{c}}$, we have $\widehat{\boldsymbol{f}} = \widehat{\boldsymbol{\tau}} + \widehat{\boldsymbol{c}}$. Hence, we obtain (2.14) and (2.15). □

2.3.2 Penalized regressions to compute $\widehat{\boldsymbol{c}}$

Concerning $\widehat{\boldsymbol{c}}$, we have the following results:

Lemma 2.8. *Consider the following penalized regressions:*

$$
\widehat{\gamma} = \arg\min_{\gamma\in\mathbb{R}^{n-2}} \|\boldsymbol{y} - \mathcal{A}_{\mathrm{r}}^{-1}\gamma\|^2 + \lambda\|\gamma\|^2 = (\mathcal{A}_{\mathrm{r}}^{-1\top}\mathcal{A}_{\mathrm{r}}^{-1} + \lambda\boldsymbol{I}_{n-2})^{-1}\mathcal{A}_{\mathrm{r}}^{-1\top}\boldsymbol{y}, \tag{2.16}
$$

$$
\widehat{\boldsymbol{\kappa}} = \arg\min_{\boldsymbol{\kappa}\in\mathbb{R}^{n-2}} \|\boldsymbol{y} - \mathcal{B}_{\mathrm{r}}^{-1}\boldsymbol{\kappa}\|^2 + \lambda\boldsymbol{\kappa}^\top\mathcal{Q}^{-1}\boldsymbol{\kappa} = (\mathcal{B}_{\mathrm{r}}^{-1\top}\mathcal{B}_{\mathrm{r}}^{-1} + \lambda\mathcal{Q}^{-1})^{-1}\mathcal{B}_{\mathrm{r}}^{-1\top}\boldsymbol{y}. \tag{2.17}
$$

Then, we have

$$
\widehat{\boldsymbol{c}} = \mathcal{A}_{\mathrm{r}}^{-1}\widehat{\gamma} = \mathcal{B}_{\mathrm{r}}^{-1}\widehat{\boldsymbol{\kappa}}. \tag{2.18}
$$

Proof of Lemma 2.8. Let $\boldsymbol{K} = [\mathcal{P}, \mathcal{A}_{\mathrm{r}}^{-1}]$. From Proposition 2.1, it follows that $\mathcal{A}\mathcal{P} = \boldsymbol{0}$, $\mathcal{A}_{\mathrm{r}}^{-1\top}\mathcal{P} = \boldsymbol{0}$, and $\boldsymbol{K}$ is nonsingular. Accordingly, given that $\boldsymbol{K}^\top\boldsymbol{K} = \mathsf{diag}(\mathcal{P}^\top\mathcal{P}, \mathcal{A}_{\mathrm{r}}^{-1\top}\mathcal{A}_{\mathrm{r}}^{-1})$ and $\mathcal{A}\boldsymbol{K} = [\mathcal{A}\mathcal{P}, \mathcal{A}\mathcal{A}_{\mathrm{r}}^{-1}] = [\boldsymbol{0}, \boldsymbol{I}_{n-2}]$, it follows that

$$
\begin{aligned}
\widehat{\boldsymbol{f}} &= \boldsymbol{K}\left(\boldsymbol{K}^\top\boldsymbol{K} + \lambda\boldsymbol{K}^\top\mathcal{A}^\top\mathcal{A}\boldsymbol{K}\right)^{-1}\boldsymbol{K}^\top\boldsymbol{y} \\
&= [\mathcal{P}, \mathcal{A}_{\mathrm{r}}^{-1}]\begin{bmatrix} (\mathcal{P}^\top\mathcal{P})^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & (\mathcal{A}_{\mathrm{r}}^{-1\top}\mathcal{A}_{\mathrm{r}}^{-1} + \lambda\boldsymbol{I}_{n-2})^{-1} \end{bmatrix}\begin{bmatrix} \mathcal{P}^\top \\ \mathcal{A}_{\mathrm{r}}^{-1\top} \end{bmatrix}\boldsymbol{y} \\
&= \mathcal{P}(\mathcal{P}^\top\mathcal{P})^{-1}\mathcal{P}^\top\boldsymbol{y} + \mathcal{A}_{\mathrm{r}}^{-1}(\mathcal{A}_{\mathrm{r}}^{-1\top}\mathcal{A}_{\mathrm{r}}^{-1} + \lambda\boldsymbol{I}_{n-2})^{-1}\mathcal{A}_{\mathrm{r}}^{-1\top}\boldsymbol{y} = \widehat{\boldsymbol{\tau}} + \mathcal{A}_{\mathrm{r}}^{-1}\widehat{\gamma},
\end{aligned}
$$

from which we have $\widehat{\boldsymbol{f}} - \widehat{\boldsymbol{\tau}} = \mathcal{A}_{\mathrm{r}}^{-1}\widehat{\gamma}$. Given $\widehat{\boldsymbol{f}} - \widehat{\boldsymbol{\tau}} = \widehat{\boldsymbol{c}}$, we thus obtain $\widehat{\boldsymbol{c}} = \mathcal{A}_{\mathrm{r}}^{-1}\widehat{\gamma}$. Similarly, we can obtain $\widehat{\boldsymbol{c}} = \mathcal{B}_{\mathrm{r}}^{-1}\widehat{\boldsymbol{\kappa}}$. □

Lemma 2.9. *$\widehat{c}$ can be calculated by the following penalized regressions:*

$$\widehat{\boldsymbol{c}} = \arg\min_{\boldsymbol{c}\in\mathbb{R}^n} \|(\boldsymbol{y}-\widehat{\boldsymbol{\tau}})-\boldsymbol{c}\|^2+\lambda\|\mathcal{A}\boldsymbol{c}\|^2= \left(\boldsymbol{I}_n+\lambda\mathcal{A}^\top\mathcal{A}\right)^{-1}(\boldsymbol{y}-\widehat{\boldsymbol{\tau}}) \tag{2.19}$$

$$= \arg\min_{\boldsymbol{c}\in\mathbb{R}^n} \|(\boldsymbol{y}-\widehat{\boldsymbol{\tau}})-\boldsymbol{c}\|^2+\lambda\boldsymbol{c}^\top\mathcal{B}^\top\mathcal{Q}^{-1}\mathcal{B}\boldsymbol{c} = \left(\boldsymbol{I}_n+\lambda\mathcal{B}^\top\mathcal{Q}^{-1}\mathcal{B}\right)^{-1}(\boldsymbol{y}-\widehat{\boldsymbol{\tau}}). \tag{2.20}$$

Proof of Lemma 2.9. Given (2.14), $\widehat{\boldsymbol{f}}=\widehat{\boldsymbol{\tau}}+\widehat{\boldsymbol{c}}$, and $\mathcal{A}\mathcal{P}=\boldsymbol{0}$, we have

$$\boldsymbol{y}=(\boldsymbol{I}_n+\lambda\mathcal{A}^\top\mathcal{A})\widehat{\boldsymbol{f}}=(\boldsymbol{I}_n+\lambda\mathcal{A}^\top\mathcal{A})(\widehat{\boldsymbol{\tau}}+\widehat{\boldsymbol{c}})=\widehat{\boldsymbol{\tau}}+(\boldsymbol{I}_n+\lambda\mathcal{A}^\top\mathcal{A})\widehat{\boldsymbol{c}},$$

which leads to (2.19). Similarly, we can obtain (2.20). □

Remark 2.2. We add some more exposition about (2.16). Let $\boldsymbol{K}=[\mathcal{P},\mathcal{A}_{\mathrm{r}}^{-1}]$ as before. In addition, let $\boldsymbol{\theta}=[\boldsymbol{\beta}^\top,\boldsymbol{\gamma}^\top]^\top\in\mathbb{R}^n$ be a vector such that $\boldsymbol{f}=\boldsymbol{K}\boldsymbol{\theta}=\mathcal{P}\boldsymbol{\beta}+\mathcal{A}_{\mathrm{r}}^{-1}\boldsymbol{\gamma}$. Then, it follows that $\mathcal{A}\boldsymbol{f}=\mathcal{A}(\mathcal{P}\boldsymbol{\beta}+\mathcal{A}_{\mathrm{r}}^{-1}\boldsymbol{\gamma})=\boldsymbol{\gamma}$. Given that $\boldsymbol{f}=\mathcal{P}\boldsymbol{\beta}+\mathcal{A}_{\mathrm{r}}^{-1}\boldsymbol{\gamma}$ and $\mathcal{A}\boldsymbol{f}=\boldsymbol{\gamma}$, the minimization problem in (2.14) can be represented as follows:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^2,\boldsymbol{\gamma}\in\mathbb{R}^{n-2}} \|\boldsymbol{y}-\mathcal{P}\boldsymbol{\beta}-\mathcal{A}_{\mathrm{r}}^{-1}\boldsymbol{\gamma}\|^2+\lambda\|\boldsymbol{\gamma}\|^2. \tag{2.21}$$

It is noteworthy that $\boldsymbol{\beta}$ is not penalized in (2.21) and $(\mathcal{A}_{\mathrm{r}}^{-1})^\top\mathcal{P}=\boldsymbol{0}$. Thus, the minimization problem (2.21) can be decomposed into (2.16) and (2.40). Moreover, (2.21) gives the best linear unbiased predictors of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ of the following linear mixed model:

$$\boldsymbol{y}=\mathcal{P}\boldsymbol{\beta}+\mathcal{A}_{\mathrm{r}}^{-1}\boldsymbol{\gamma}+\boldsymbol{u},\quad [\boldsymbol{u}^\top,\boldsymbol{\gamma}^\top]^\top\sim N\left(\boldsymbol{0},\mathsf{diag}(\sigma_u^2\boldsymbol{I}_n,\sigma_v^2\boldsymbol{I}_{n-2})\right), \tag{2.22}$$

where $\lambda=\sigma_u^2/\sigma_v^2$.

Remark 2.3. By using $\boldsymbol{C}_{\mathrm{r}}^{-1}$, Verbyla et al. (1999) derived the following expressions in our notations:

$$\widehat{\boldsymbol{f}}=\widehat{\boldsymbol{\tau}}+\boldsymbol{C}_{\mathrm{r}}^{-1}\widehat{\boldsymbol{\kappa}},\quad \widehat{\boldsymbol{\kappa}}=(\boldsymbol{C}_{\mathrm{r}}^{-1\top}\boldsymbol{C}_{\mathrm{r}}^{-1}+\lambda\boldsymbol{R}^{-1})^{-1}\boldsymbol{C}_{\mathrm{r}}^{-1\top}\boldsymbol{y}. \tag{2.23}$$

Here, we make the following remarks on (2.23). (i) First, $\widehat{\boldsymbol{\kappa}}$ is the solution of the following penalized regression:

$$\min_{\boldsymbol{\kappa}\in\mathbb{R}^{n-2}} \|\boldsymbol{y}-\boldsymbol{C}_{\mathrm{r}}^{-1}\boldsymbol{\kappa}\|^2+\lambda\boldsymbol{\kappa}^\top\boldsymbol{R}^{-1}\boldsymbol{\kappa}. \tag{2.24}$$

(ii) Moreover, (2.23) is a special case of $\widehat{\boldsymbol{c}}=\mathcal{B}_{\mathrm{r}}^{-1}\widehat{\boldsymbol{\kappa}}$ in (2.18).

2.3.3 Penalized regressions to compute $\widehat{\boldsymbol{u}}$

Concerning $\widehat{\boldsymbol{u}}$, we have the following results:

Lemma 2.10. *Consider the following penalized regressions:*

$$\widehat{\boldsymbol{\eta}}=\arg\min_{\boldsymbol{\eta}\in\mathbb{R}^{n-2}} \|\boldsymbol{y}-\mathcal{A}^\top\boldsymbol{\eta}\|^2+\lambda^{-1}\|\boldsymbol{\eta}\|^2=(\mathcal{A}\mathcal{A}^\top+\lambda^{-1}\boldsymbol{I}_{n-2})^{-1}\mathcal{A}\boldsymbol{y}, \tag{2.25}$$

$$\widehat{\boldsymbol{v}}=\arg\min_{\boldsymbol{v}\in\mathbb{R}^{n-2}} \|\boldsymbol{y}-\mathcal{B}^\top\boldsymbol{v}\|^2+\lambda^{-1}\boldsymbol{v}^\top\mathcal{Q}\boldsymbol{v}=(\mathcal{B}\mathcal{B}^\top+\lambda^{-1}\mathcal{Q})^{-1}\mathcal{B}\boldsymbol{y}. \tag{2.26}$$

Then, it follows that

$$\widehat{\boldsymbol{u}}=\mathcal{A}^\top\widehat{\boldsymbol{\eta}}=\mathcal{B}^\top\widehat{\boldsymbol{v}}. \tag{2.27}$$

Proof of Lemma 2.10. Applying the matrix inversion lemma to $\left(\boldsymbol{I}_n+\lambda\mathcal{A}^\top\mathcal{A}\right)^{-1}$, we have

$$\left(\boldsymbol{I}_n+\lambda\mathcal{A}^\top\mathcal{A}\right)^{-1}=\boldsymbol{I}_n-\mathcal{A}^\top(\mathcal{A}\mathcal{A}^\top+\lambda^{-1}\boldsymbol{I}_{n-2})^{-1}\mathcal{A}. \tag{2.28}$$

Postmultiplying (2.28) by $\boldsymbol{y}$ yields $\boldsymbol{y}-\widehat{\boldsymbol{f}}=\mathcal{A}^\top\widehat{\boldsymbol{\eta}}$. Given $\boldsymbol{y}-\widehat{\boldsymbol{f}}=\widehat{\boldsymbol{u}}$, we thus have $\widehat{\boldsymbol{u}}=\mathcal{A}^\top\widehat{\boldsymbol{\eta}}$. Similarly, we can obtain $\widehat{\boldsymbol{u}}=\mathcal{B}^\top\widehat{\boldsymbol{v}}$. □

Lemma 2.11. *$\widehat{\boldsymbol{u}}$ can be calculated by the following penalized regressions:*

$$\begin{aligned}\widehat{\boldsymbol{u}}&=\arg\min_{\boldsymbol{u}\in\mathbb{R}^n} \|(\boldsymbol{y}-\widehat{\boldsymbol{\tau}})-\boldsymbol{u}\|^2+\lambda^{-1}\|\mathcal{A}_{\mathrm{r}}^{-1\top}\boldsymbol{u}\|^2\\&=\left(\boldsymbol{I}_n+\lambda^{-1}\mathcal{A}_{\mathrm{r}}^{-1}\mathcal{A}_{\mathrm{r}}^{-1\top}\right)^{-1}(\boldsymbol{y}-\widehat{\boldsymbol{\tau}})\end{aligned} \tag{2.29}$$

and

$$\begin{aligned}\widehat{\boldsymbol{u}} &= \arg\min_{\boldsymbol{u}\in\mathbb{R}^n} \|(\boldsymbol{y}-\widehat{\boldsymbol{\tau}})-\boldsymbol{u}\|^2+\lambda^{-1}\boldsymbol{u}^\top\mathcal{B}_{\mathrm{r}}^{-1}\mathcal{Q}\mathcal{B}_{\mathrm{r}}^{-1\top}\boldsymbol{u}\\ &= \left(\boldsymbol{I}_n+\lambda^{-1}\mathcal{B}_{\mathrm{r}}^{-1}\mathcal{Q}\mathcal{B}_{\mathrm{r}}^{-1\top}\right)^{-1}(\boldsymbol{y}-\widehat{\boldsymbol{\tau}}). \end{aligned} \tag{2.30}$$

Proof of Lemma 2.11. Given (2.34), $\widehat{\boldsymbol{g}}=\widehat{\boldsymbol{\tau}}+\widehat{\boldsymbol{u}}$, and $\mathcal{A}_{\mathrm{r}}^{-1\top}\mathcal{P}=\mathbf{0}$, we have

$$\begin{aligned}\boldsymbol{y} &= (\boldsymbol{I}_n+\lambda^{-1}\mathcal{A}_{\mathrm{r}}^{-1}\mathcal{A}_{\mathrm{r}}^{-1\top})\widehat{\boldsymbol{g}} = (\boldsymbol{I}_n+\lambda^{-1}\mathcal{A}_{\mathrm{r}}^{-1}\mathcal{A}_{\mathrm{r}}^{-1\top})(\widehat{\boldsymbol{\tau}}+\widehat{\boldsymbol{u}})\\ &= \widehat{\boldsymbol{\tau}}+(\boldsymbol{I}_n+\lambda^{-1}\mathcal{A}_{\mathrm{r}}^{-1}\mathcal{A}_{\mathrm{r}}^{-1\top})\widehat{\boldsymbol{u}},\end{aligned}$$

which leads to (2.29). Similarly, we can obtain (2.30). □

Remark 2.4. In Reinsch (1967) and Green and Silverman (1994, p. 20), there are equations expressed in our notation as follows:

$$(\boldsymbol{R}+\lambda\boldsymbol{C}\boldsymbol{C}^\top)\boldsymbol{\phi}=\boldsymbol{C}\boldsymbol{y}, \quad \widehat{\boldsymbol{f}}=\boldsymbol{y}-\lambda\boldsymbol{C}^\top\boldsymbol{\phi}. \tag{2.31}$$

Here, we make the following remarks on (2.31). (i) First, these lead to a penalized least squares problem. Given that $\widehat{\boldsymbol{u}}=\boldsymbol{y}-\widehat{\boldsymbol{f}}$, removing $\boldsymbol{\phi}$ from the above equations leads to

$$\begin{aligned}\widehat{\boldsymbol{u}} &= \boldsymbol{y}-\widehat{\boldsymbol{f}}=\lambda\boldsymbol{C}^\top(\boldsymbol{R}+\lambda\boldsymbol{C}\boldsymbol{C}^\top)^{-1}\boldsymbol{C}\boldsymbol{y}\\ &= \boldsymbol{C}^\top(\boldsymbol{C}\boldsymbol{C}^\top+\lambda^{-1}\boldsymbol{R})^{-1}\boldsymbol{C}\boldsymbol{y}=\boldsymbol{C}^\top\widehat{\boldsymbol{v}},\end{aligned} \tag{2.32}$$

where

$$\widehat{\boldsymbol{v}}=\arg\min_{\boldsymbol{v}\in\mathbb{R}^{n-2}}\|\boldsymbol{y}-\boldsymbol{C}^\top\boldsymbol{v}\|^2+\lambda^{-1}\boldsymbol{v}^\top\boldsymbol{R}\boldsymbol{v}. \tag{2.33}$$

(ii) Moreover, (2.32) is a special case of $\widehat{\boldsymbol{u}}=\mathcal{B}^\top\widehat{\boldsymbol{v}}$ in (2.27).

2.3.4 Penalized regression to compute $\widehat{\boldsymbol{\tau}}+\widehat{\boldsymbol{u}}$

Concerning $\widehat{\boldsymbol{\tau}}+\widehat{\boldsymbol{u}}$, we have the following results:

Lemma 2.12. *Let $\widehat{\boldsymbol{g}} = \widehat{\boldsymbol{\tau}} + \widehat{\boldsymbol{u}}$. Then, it follows that*

$$\widehat{\boldsymbol{g}} = \arg\min_{\boldsymbol{g}\in\mathbb{R}^n} \|\boldsymbol{y}-\boldsymbol{g}\|^2 + \lambda^{-1}\|\mathcal{A}_{\mathrm{r}}^{-1\top}\boldsymbol{g}\|^2 = \left(\boldsymbol{I}_n + \lambda^{-1}\mathcal{A}_{\mathrm{r}}^{-1}\mathcal{A}_{\mathrm{r}}^{-1\top}\right)^{-1}\boldsymbol{y} \tag{2.34}$$

$$= \arg\min_{\boldsymbol{g}\in\mathbb{R}^n} \|\boldsymbol{y}-\boldsymbol{g}\|^2 + \lambda^{-1}\boldsymbol{g}^\top\mathcal{B}_{\mathrm{r}}^{-1}\mathcal{Q}\mathcal{B}_{\mathrm{r}}^{-1\top}\boldsymbol{g} = \left(\boldsymbol{I}_n + \lambda^{-1}\mathcal{B}_{\mathrm{r}}^{-1}\mathcal{Q}\mathcal{B}_{\mathrm{r}}^{-1\top}\right)^{-1}\boldsymbol{y}. \tag{2.35}$$

Proof of Lemma 2.12. Let $\boldsymbol{J} = [\mathcal{P}, \mathcal{A}^\top]$. From Proposition 2.1, it follows that $\mathcal{A}\mathcal{P} = \boldsymbol{0}$, $\mathcal{A}_{\mathrm{r}}^{-1\top}\mathcal{P} = \boldsymbol{0}$, and $\boldsymbol{J}$ is nonsingular. Accordingly, given that $\boldsymbol{J}^\top\boldsymbol{J} = \mathsf{diag}(\mathcal{P}^\top\mathcal{P}, \mathcal{A}\mathcal{A}^\top)$ and $\mathcal{A}_{\mathrm{r}}^{-1\top}\boldsymbol{J} = [\mathcal{A}_{\mathrm{r}}^{-1\top}\mathcal{P}, \mathcal{A}_{\mathrm{r}}^{-1\top}\mathcal{A}^\top] = [\boldsymbol{0}, \boldsymbol{I}_{n-2}]$, it follows that

$$\begin{aligned}
&\left(\boldsymbol{I}_n + \lambda^{-1}\mathcal{A}_{\mathrm{r}}^{-1}\mathcal{A}_{\mathrm{r}}^{-1\top}\right)^{-1}\boldsymbol{y} \\
&\quad= \boldsymbol{J}\left(\boldsymbol{J}^\top\boldsymbol{J} + \lambda^{-1}\boldsymbol{J}^\top\mathcal{A}_{\mathrm{r}}^{-1}\mathcal{A}_{\mathrm{r}}^{-1\top}\boldsymbol{J}\right)^{-1}\boldsymbol{J}^\top\boldsymbol{y} \\
&\quad= [\mathcal{P}, \mathcal{A}^\top]\begin{bmatrix} (\mathcal{P}^\top\mathcal{P})^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & (\mathcal{A}\mathcal{A}^\top + \lambda^{-1}\boldsymbol{I}_{n-2})^{-1} \end{bmatrix}\begin{bmatrix} \mathcal{P}^\top \\ \mathcal{A} \end{bmatrix}\boldsymbol{y} \\
&\quad= \mathcal{P}(\mathcal{P}^\top\mathcal{P})^{-1}\mathcal{P}^\top\boldsymbol{y} + \mathcal{A}^\top(\mathcal{A}\mathcal{A}^\top + \lambda^{-1}\boldsymbol{I}_{n-2})^{-1}\mathcal{A}\boldsymbol{y} = \widehat{\boldsymbol{\tau}} + \mathcal{A}^\top\widehat{\boldsymbol{\eta}}.
\end{aligned}$$

Given $\widehat{\boldsymbol{u}} = \mathcal{A}^\top\widehat{\boldsymbol{\eta}}$, we obtain (2.34). Similarly, we can obtain (2.35). □

Remark 2.5. Similarly to Remark 2.3.2, we add some more exposition about (2.25). Let $\boldsymbol{\xi} = [\boldsymbol{\beta}^\top, \boldsymbol{\eta}^\top]^\top \in \mathbb{R}^n$ be such that $\boldsymbol{g} = \boldsymbol{J}\boldsymbol{\xi} = \mathcal{P}\boldsymbol{\beta} + \mathcal{A}^\top\boldsymbol{\eta}$. As stated, $\mathcal{A}_{\mathrm{r}}^{-1\top}\boldsymbol{J} = [\boldsymbol{0}, \boldsymbol{I}_{n-2}]$. Then, it follows that

$$\mathcal{A}_{\mathrm{r}}^{-1\top}\boldsymbol{g} = \mathcal{A}_{\mathrm{r}}^{-1\top}\boldsymbol{J}\boldsymbol{\xi} = \boldsymbol{\eta}. \tag{2.36}$$

Given $\boldsymbol{g} = \mathcal{P}\boldsymbol{\beta} + \mathcal{A}^\top\boldsymbol{\eta}$ and $\mathcal{A}_{\mathrm{r}}^{-1\top}\boldsymbol{g} = \boldsymbol{\eta}$, the minimization problem (2.34) can be represented as follows:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^2, \boldsymbol{\eta}\in\mathbb{R}^{n-2}} \|\boldsymbol{y} - \mathcal{P}\boldsymbol{\beta} - \mathcal{A}^\top\boldsymbol{\eta}\|^2 + \lambda^{-1}\|\boldsymbol{\eta}\|^2. \tag{2.37}$$

Again, it is noteworthy that $\boldsymbol{\beta}$ is not penalized in (2.37). Moreover, it follows that $(\mathcal{A}^\top)^\top\mathcal{P} = \mathcal{A}\mathcal{P} = \boldsymbol{0}$. Thus, the minimization problem (2.37) can be decomposed into (2.25) and (2.40).

2.3.5 Ordinary regressions to compute $\widehat{\boldsymbol{c}}+\widehat{\boldsymbol{u}}$ and $\widehat{\boldsymbol{\tau}}$

Concerning $\widehat{\boldsymbol{c}}+\widehat{\boldsymbol{u}}$ and $\widehat{\boldsymbol{\tau}}$, we have the following results:

Lemma 2.13. *(i) Let $\widehat{\boldsymbol{h}}=\mathcal{D}^\top\widehat{\boldsymbol{\alpha}}$, where*

$$\widehat{\boldsymbol{\alpha}}=\arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^{n-2}}\|\boldsymbol{y}-\mathcal{D}^\top\boldsymbol{\alpha}\|^2=(\mathcal{D}\mathcal{D}^\top)^{-1}\mathcal{D}\boldsymbol{y}. \tag{2.38}$$

Then, it follows that

$$\widehat{\boldsymbol{c}}+\widehat{\boldsymbol{u}}=\widehat{\boldsymbol{h}}. \tag{2.39}$$

(ii) It follows that $\widehat{\boldsymbol{\tau}}=\mathcal{P}\widehat{\boldsymbol{\beta}}$, where

$$\widehat{\boldsymbol{\beta}}=\arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{2}}\|\boldsymbol{y}-\mathcal{P}\boldsymbol{\beta}\|^2=(\mathcal{P}^\top\mathcal{P})^{-1}\mathcal{P}^\top\boldsymbol{y}. \tag{2.40}$$

Proof of Lemma 2.13. Given Proposition 2.1, both results are easily obtainable. For example, the former result can be proved as follows:

$$\widehat{\boldsymbol{h}}=\mathcal{D}^\top\widehat{\boldsymbol{\alpha}}=\boldsymbol{P}_{\mathcal{D}^\top}\boldsymbol{y}=\boldsymbol{Q}_{\mathcal{P}}\boldsymbol{y}=\boldsymbol{y}-\widehat{\boldsymbol{\tau}}=\widehat{\boldsymbol{c}}+\widehat{\boldsymbol{u}}.$$

□

Remark 2.6. From Proposition 2.1, we also have $\widehat{\boldsymbol{h}}(=\widehat{\boldsymbol{c}}+\widehat{\boldsymbol{u}})=\mathcal{D}_{\mathrm{r}}^{-1}\widehat{\boldsymbol{\rho}}$, where

$$\widehat{\boldsymbol{\rho}}=\arg\min_{\boldsymbol{\rho}\in\mathbb{R}^{n-2}}\|\boldsymbol{y}-\mathcal{D}_{\mathrm{r}}^{-1}\boldsymbol{\rho}\|^2=(\mathcal{D}_{\mathrm{r}}^{-1\top}\mathcal{D}_{\mathrm{r}}^{-1})^{-1}\mathcal{D}_{\mathrm{r}}^{-1\top}\boldsymbol{y}. \tag{2.41}$$

2.3.6 Principle of duality in the penalized regressions

See the second columns of Tables 1–2. In the columns, the penalized regressions shown above are arranged in pairs that mirror one another. We reveal a principle of duality in the penalized regressions. As stated in Section 1, (D1) is obtainable by replacing $\mathcal{A}^\top,\lambda$ in (P1) by $\mathcal{A}_{\mathrm{r}}^{-1},\lambda^{-1}$, respectively. Likewise, for example, (D6) in Table 2 is obtainable by replacing $\mathcal{B}^\top$, $\mathcal{Q},\lambda^{-1}$ in (P6) by $\mathcal{B}_{\mathrm{r}}^{-1}$, $\mathcal{Q}^{-1},\lambda$, respectively. In Tables 1–2, we may observe five other pairs of regressions that are duals of each other. From the seven dual pairs shown in Tables 1–2, we observe that the following principle exists:

Proposition 2.2 (Principle of duality)**.** *The regressions labeled with the letter D in Tables 1–2, e.g., (D1), are obtainable by replacing each occurrence of* $\mathcal{A}^\top, \mathcal{B}^\top, \mathcal{D}^\top, \mathcal{Q}, \mathcal{Q}^{-1}, \lambda, \lambda^{-1}$ *in the regressions labeled with the letter P, e.g., (P1), by* $\mathcal{A}_{\mathrm{r}}^{-1}, \mathcal{B}_{\mathrm{r}}^{-1}, \mathcal{D}_{\mathrm{r}}^{-1}, \mathcal{Q}^{-1}, \mathcal{Q}, \lambda^{-1}, \lambda$*, respectively.*

2.4 Results that are obtainable from the regressions

In this section, we show how the regressions listed in the previous section are of use to obtain a deeper understanding of the fitting a cubic smoothing spline. Before proceeding, recall $\widehat{\boldsymbol{f}} = \widehat{\boldsymbol{\tau}} + \widehat{\boldsymbol{c}}$ and so on.

First, given that (2.16) is a ridge regression, it immediately follows that $\lim_{\lambda\to\infty} \widehat{\boldsymbol{\gamma}} = \boldsymbol{0}$, which leads to $\lim_{\lambda\to\infty} \widehat{\boldsymbol{c}} = \mathcal{A}_{\mathrm{r}}^{-1} \lim_{\lambda\to\infty} \widehat{\boldsymbol{\gamma}} = \boldsymbol{0}$ and at the same time we have

$$\lim_{\lambda\to\infty} \widehat{\boldsymbol{f}} = \widehat{\boldsymbol{\tau}} + \lim_{\lambda\to\infty} \widehat{\boldsymbol{c}} = \widehat{\boldsymbol{\tau}}, \tag{2.42}$$

$$\lim_{\lambda\to\infty} \widehat{\boldsymbol{u}} = \boldsymbol{y} - \widehat{\boldsymbol{\tau}} - \lim_{\lambda\to\infty} \widehat{\boldsymbol{c}} = \boldsymbol{y} - \widehat{\boldsymbol{\tau}}, \tag{2.43}$$

$$\lim_{\lambda\to\infty} \widehat{\boldsymbol{g}} = \widehat{\boldsymbol{\tau}} + \lim_{\lambda\to\infty} \widehat{\boldsymbol{u}} = \widehat{\boldsymbol{\tau}} + (\boldsymbol{y} - \widehat{\boldsymbol{\tau}}) = \boldsymbol{y}. \tag{2.44}$$

Second, (2.25) is again a ridge regression, we have $\lim_{\lambda\to 0} \widehat{\boldsymbol{\eta}} = \boldsymbol{0}$, which yields $\lim_{\lambda\to 0} \widehat{\boldsymbol{u}} = \mathcal{A}^\top \lim_{\lambda\to 0} \widehat{\boldsymbol{\eta}} = \boldsymbol{0}$ and accordingly we obtain

$$\lim_{\lambda\to 0} \widehat{\boldsymbol{f}} = \boldsymbol{y} - \lim_{\lambda\to 0} \widehat{\boldsymbol{u}} = \boldsymbol{y}, \tag{2.45}$$

$$\lim_{\lambda\to 0} \widehat{\boldsymbol{c}} = \boldsymbol{y} - \widehat{\boldsymbol{\tau}} - \lim_{\lambda\to 0} \widehat{\boldsymbol{u}} = \boldsymbol{y} - \widehat{\boldsymbol{\tau}}, \tag{2.46}$$

$$\lim_{\lambda\to 0} \widehat{\boldsymbol{g}} = \widehat{\boldsymbol{\tau}} + \lim_{\lambda\to 0} \widehat{\boldsymbol{u}} = \widehat{\boldsymbol{\tau}}. \tag{2.47}$$

Third, from (2.19) and $\widehat{\boldsymbol{u}} = \boldsymbol{y} - \widehat{\boldsymbol{\tau}} - \widehat{\boldsymbol{c}}$, we have

$$\widehat{\boldsymbol{c}} = \left(\boldsymbol{I}_n + \lambda \mathcal{A}^\top \mathcal{A}\right)^{-1} (\boldsymbol{y} - \widehat{\boldsymbol{\tau}}), \tag{2.48}$$

$$\widehat{\boldsymbol{u}} = \left\{\boldsymbol{I}_n - (\boldsymbol{I}_n + \lambda \mathcal{A}^\top \mathcal{A})^{-1}\right\} (\boldsymbol{y} - \widehat{\boldsymbol{\tau}}). \tag{2.49}$$

Thus, $\widehat{\boldsymbol{f}}$ can be represented as

$$\widehat{\boldsymbol{f}} = \widehat{\boldsymbol{\tau}} + \left(\boldsymbol{I}_n + \lambda \mathcal{A}^\top \mathcal{A}\right)^{-1} (\boldsymbol{y} - \widehat{\boldsymbol{\tau}}). \tag{2.50}$$

Here, we remark that, given that $\left(\boldsymbol{I}_n + \lambda \mathcal{A}^\top \mathcal{A}\right)^{-1}$ is a smoother matrix, the second term on the right-hand side of (2.50) represents a low-frequency part of $\boldsymbol{y} - \widehat{\boldsymbol{\tau}}$. In addition, from (2.49), $\widehat{\boldsymbol{u}}$ represents a high-frequency part of $\boldsymbol{y} - \widehat{\boldsymbol{\tau}}$. Thus, $\widehat{\boldsymbol{c}}$ is generally smoother than $\widehat{\boldsymbol{u}}$.

Fourth, given $\mathcal{A}\mathcal{P} = \boldsymbol{0}$, $\mathcal{A}_{\mathrm{r}}^{-1\top}\mathcal{P} = \boldsymbol{0}$, $\widehat{\boldsymbol{c}} = \mathcal{A}_{\mathrm{r}}^{-1}\widehat{\boldsymbol{\gamma}}$, and $\widehat{\boldsymbol{u}} = \mathcal{A}^\top \widehat{\boldsymbol{\eta}}$, we have

$$\widehat{\boldsymbol{\zeta}}^\top \widehat{\boldsymbol{\tau}} = 0, \quad \widehat{\boldsymbol{\zeta}} = \widehat{\boldsymbol{c}}, \widehat{\boldsymbol{u}}, \widehat{\boldsymbol{h}}. \tag{2.51}$$

Fifth, given $\mathcal{A}\mathcal{P} = \boldsymbol{0}$, $\mathcal{A}_{\mathrm{r}}^{-1\top}\mathcal{P} = \boldsymbol{0}$, (2.28), and

$$\left(\boldsymbol{I}_n + \lambda^{-1}\mathcal{A}_{\mathrm{r}}^{-1}\mathcal{A}_{\mathrm{r}}^{-1\top}\right)^{-1} = \boldsymbol{I}_n - \mathcal{A}_{\mathrm{r}}^{-1}(\mathcal{A}_{\mathrm{r}}^{-1\top}\mathcal{A}_{\mathrm{r}}^{-1} + \lambda \boldsymbol{I}_{n-2})^{-1}\mathcal{A}_{\mathrm{r}}^{-1\top}, \tag{2.52}$$

if $\boldsymbol{y} \in \mathcal{S}(\mathcal{P})$, or in other words, if $\boldsymbol{y} = \mathcal{P}\boldsymbol{\psi}$, then we have

$$\widehat{\boldsymbol{\tau}} = \boldsymbol{y}, \quad \widehat{\boldsymbol{f}} = \boldsymbol{y}, \quad \widehat{\boldsymbol{g}} = \boldsymbol{y}, \quad \widehat{\boldsymbol{c}} = \boldsymbol{0}, \quad \widehat{\boldsymbol{u}} = \boldsymbol{0}, \quad \widehat{\boldsymbol{h}} = \boldsymbol{0}. \tag{2.53}$$

Sixth, given $\boldsymbol{\iota}_n \in \mathcal{S}(\mathcal{P})$, we have

$$\boldsymbol{P}_{\mathcal{P}}\boldsymbol{\iota}_n = \boldsymbol{\iota}_n, \tag{2.54}$$

$$(\boldsymbol{I}_n + \lambda \mathcal{A}^\top \mathcal{A})^{-1}\boldsymbol{\iota}_n = \boldsymbol{\iota}_n, \tag{2.55}$$

$$(\boldsymbol{I}_n + \lambda^{-1}\mathcal{A}_{\mathrm{r}}^{-1}\mathcal{A}_{\mathrm{r}}^{-1\top})^{-1}\boldsymbol{\iota}_n = \boldsymbol{\iota}_n, \tag{2.56}$$

$$\mathcal{A}_{\mathrm{r}}^{-1}(\mathcal{A}_{\mathrm{r}}^{-1\top}\mathcal{A}_{\mathrm{r}}^{-1} + \lambda \boldsymbol{I}_{n-2})^{-1}\mathcal{A}_{\mathrm{r}}^{-1\top}\boldsymbol{\iota}_n = \boldsymbol{0}, \tag{2.57}$$

$$\mathcal{A}^\top(\mathcal{A}\mathcal{A}^\top + \lambda^{-1}\boldsymbol{I}_{n-2})^{-1}\mathcal{A}\boldsymbol{\iota}_n = \boldsymbol{0}, \tag{2.58}$$

$$\boldsymbol{P}_{\mathcal{D}^\top}\boldsymbol{\iota}_n = \boldsymbol{0}. \tag{2.59}$$

Note that $(\boldsymbol{I}_n + \lambda \mathcal{A}^\top \mathcal{A})^{-1}\boldsymbol{\iota}_n = \boldsymbol{\iota}_n$, for example, indicates that the sum of the entries in each row of the hat matrix of $\widehat{\boldsymbol{f}}$ equals unity.

Seventh, given (2.54)–(2.59), we have

$$\frac{1}{n}\boldsymbol{\iota}_n^\top \widehat{\boldsymbol{\zeta}} = \bar{y}, \quad \widehat{\boldsymbol{\zeta}} = \widehat{\boldsymbol{\tau}}, \widehat{\boldsymbol{f}}, \widehat{\boldsymbol{g}}, \tag{2.60}$$

$$\frac{1}{n}\boldsymbol{\iota}_n^\top \widehat{\boldsymbol{\zeta}} = 0, \quad \widehat{\boldsymbol{\zeta}} = \widehat{\boldsymbol{c}}, \widehat{\boldsymbol{u}}, \widehat{\boldsymbol{h}}, \tag{2.61}$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$. $\frac{1}{n}\boldsymbol{\iota}_n^\top \widehat{\boldsymbol{f}} = \bar{y}$, for example, shows that $\frac{1}{n}\sum_{i=1}^n \widehat{f}_i = \bar{y}$.

2.5 Illustrations of some results

In this section, we illustrate some of the results in the previous sections by a real data example.

Panel A of Figure 2.1 shows a scatter plot of the log of seasonally adjusted Japanese real gross domestic product (GDP) over the sample period 1994:Q1 to 2020:Q2 (and accordingly, $n = 106$). We obtained the data from the website of Japanese Cabinet office. The solid line in the panel plots $(x_i, \widehat{\tau}_i)$ for $i = 1, \dots, n$, where $\widehat{\boldsymbol{\tau}} = [\widehat{\tau}_1, \dots, \widehat{\tau}_n]^\top$ in (2.4) and $n = 106$. Panel B of Figure 2.1 depicts a scatter plot of $(x_i, y_i - \widehat{\tau}_i)$ for $i = 1, \dots, n$. The solid line in the panel plots $(x_i, \widehat{c}_i)$ for $i = 1, \dots, n$, where $\widehat{\boldsymbol{c}} = [\widehat{c}_1, \dots, \widehat{c}_n]^\top$ is calculated by (2.18) with $\lambda = 10^3$. The solid line in Panel C denotes $(x_i, \widehat{f}_i)$, where $\widehat{\boldsymbol{f}} = [\widehat{f}_1, \dots, \widehat{f}_n]^\top$ is calculated by (2.14) with $\lambda = 10^3$. Panel D illustrates a scatter plot of $(x_i, y_i - \widehat{\tau}_i)$ for $i = 1, \dots, n$. The solid line in the panel plots $(x_i, \widehat{u}_i)$ for $i = 1, \dots, n$, where $\widehat{\boldsymbol{u}} = [\widehat{u}_1, \dots, \widehat{u}_n]^\top$ is calculated by (2.27) with $\lambda = 10^3$. Figures 2.2, 2.3, and 2.4 correspond to the cases such that $\lambda = 10^5, 10^{10}, 10^{-10}$, respectively.

Recall that concerning $\boldsymbol{y}$, $\widehat{\boldsymbol{\tau}}$, $\widehat{\boldsymbol{c}}$, $\widehat{\boldsymbol{f}}$, and $\widehat{\boldsymbol{u}}$, the following equations hold:

$$\widehat{\boldsymbol{\tau}} + \widehat{\boldsymbol{c}} = \widehat{\boldsymbol{f}}, \quad \widehat{\boldsymbol{c}} + \widehat{\boldsymbol{u}} = \boldsymbol{y} - \widehat{\boldsymbol{\tau}}, \quad \lim_{\lambda\to\infty} \widehat{\boldsymbol{c}} = \boldsymbol{0}, \quad \lim_{\lambda\to\infty} \widehat{\boldsymbol{f}} = \widehat{\boldsymbol{\tau}},$$
$$\lim_{\lambda\to\infty} \widehat{\boldsymbol{u}} = \boldsymbol{y} - \widehat{\boldsymbol{\tau}}, \quad \lim_{\lambda\to 0} \widehat{\boldsymbol{c}} = \boldsymbol{y} - \widehat{\boldsymbol{\tau}}, \quad \lim_{\lambda\to 0} \widehat{\boldsymbol{f}} = \boldsymbol{y}, \quad \lim_{\lambda\to 0} \widehat{\boldsymbol{u}} = \boldsymbol{0}.$$

From Figures 2.1–2.4, we can observe that these theoretical results are well illustrated in these figures. For example, from Panel D in Figure 2.4, we can observe that $\widehat{\boldsymbol{u}}$ almost equals $\boldsymbol{0}$ when $\lambda = 10^{-10}$.

2.6 The cases such that the other right-inverse matrices are used

In this section, we illustrate what happens if the other right-inverse matrices are used.

Let $\boldsymbol{M} \in \mathbb{R}^{m\times n}$ be of full row rank. Recall that in this chapter $\boldsymbol{M}_{\mathrm{r}}^{-1}$ denotes $\boldsymbol{M}^\top(\boldsymbol{M}\boldsymbol{M}^\top)^{-1}$, which is a right-inverse matrix of a full-row-rank matrix $\boldsymbol{M} \in \mathbb{R}^{m\times n}$. Define a set of matrices

$$\Gamma_M = \{\boldsymbol{\Xi} \in \mathbb{R}^{n\times m} : \boldsymbol{M}\boldsymbol{\Xi} = \boldsymbol{I}_m\}.$$

Γ_M denotes the set of right-inverse matrices of $\boldsymbol{M}$ and accordingly $\boldsymbol{M}_{\mathrm{r}}^{-1}$ belongs to Γ_M.

Lemma 2.14. *$\boldsymbol{N} = \boldsymbol{M}_{\mathrm{r}}^{-1}$ if and only if $\boldsymbol{N} \in \Gamma_M$ and $\mathcal{S}(\boldsymbol{N}) = \mathcal{S}(\boldsymbol{M}^\top)$.*

Proof of Lemma 2.14. It is clear that if $\boldsymbol{N} = \boldsymbol{M}_{\mathrm{r}}^{-1}$, then $\boldsymbol{N} \in \Gamma_M$ and $\mathcal{S}(\boldsymbol{N}) = \mathcal{S}(\boldsymbol{M}^\top)$. Conversely, suppose that $\boldsymbol{N} \in \Gamma_M$ and $\mathcal{S}(\boldsymbol{N}) = \mathcal{S}(\boldsymbol{M}^\top)$. Then, $\boldsymbol{M}\boldsymbol{N} = \boldsymbol{I}_m$ and there exists a nonsingular matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{m\times m}$ such that $\boldsymbol{N} = \boldsymbol{M}^\top\boldsymbol{\Sigma}$. By removing $\boldsymbol{N}$ from these equations, we have $\boldsymbol{\Sigma} = (\boldsymbol{M}\boldsymbol{M}^\top)^{-1}$, which leads to $\boldsymbol{N} = \boldsymbol{M}^\top(\boldsymbol{M}\boldsymbol{M}^\top)^{-1} = \boldsymbol{M}_{\mathrm{r}}^{-1}$. ☐

From Lemma 2.14, if $\boldsymbol{N} \neq \boldsymbol{M}_{\mathrm{r}}^{-1}$, then $\boldsymbol{N} \notin \Gamma_M$ or $\mathcal{S}(\boldsymbol{N}) \neq \mathcal{S}(\boldsymbol{M}^\top)$. Accordingly, we have the following result:

Proposition 2.3. *If $\boldsymbol{N} \in \Gamma_M \backslash \{\boldsymbol{M}_{\mathrm{r}}^{-1}\}$, then $\mathcal{S}(\boldsymbol{N}) \neq \mathcal{S}(\boldsymbol{M}^\top)$.*

Based on the result, we illustrate what happens if the other right-inverse matrices are used. We give an example. Let $\boldsymbol{Z} \in \Gamma_D \backslash \{\boldsymbol{D}_{\mathrm{r}}^{-1}\}$. Then, from Proposition 2.3 and Lemma 2.3, it follows that $\mathcal{S}(\boldsymbol{Z}) \neq \mathcal{S}(\boldsymbol{D}_{\mathrm{r}}^{-1}) = \mathcal{S}^\perp(\boldsymbol{\Pi})$. Accordingly, letting $\boldsymbol{L} = [\boldsymbol{\Pi}, \boldsymbol{Z}]$, it follows that $\boldsymbol{Z}^\top\boldsymbol{\Pi} \neq \boldsymbol{0}$ and $\boldsymbol{D}\boldsymbol{L} = [\boldsymbol{D}\boldsymbol{\Pi}, \boldsymbol{D}\boldsymbol{Z}] = [\boldsymbol{0}, \boldsymbol{I}_{n-2}]$. In addition, given that $\boldsymbol{D}\boldsymbol{\Pi} = \boldsymbol{0}$, $\boldsymbol{D}\boldsymbol{Z} = \boldsymbol{I}_{n-2}$, and $\boldsymbol{\Pi}$ is of full column rank, $\boldsymbol{L}$ is nonsingular. Thus, from e.g., Yamada (2017), we have

$$\widehat{\boldsymbol{f}} = \boldsymbol{L}(\boldsymbol{L}^\top\boldsymbol{L} + \lambda\boldsymbol{L}^\top\boldsymbol{D}^\top\boldsymbol{D}\boldsymbol{L})^{-1}\boldsymbol{L}^\top\boldsymbol{y} = \boldsymbol{\Pi}\widehat{\boldsymbol{\pi}} + \boldsymbol{Z}\widehat{\boldsymbol{\varepsilon}}, \tag{2.62}$$

where

$$\widehat{\boldsymbol{\pi}} = \arg\min_{\boldsymbol{\pi}\in\mathbb{R}^2} \|(\boldsymbol{y} - \boldsymbol{Z}\widehat{\boldsymbol{\varepsilon}}) - \boldsymbol{\Pi}\boldsymbol{\pi}\|^2 = (\boldsymbol{\Pi}^\top\boldsymbol{\Pi})^{-1}\boldsymbol{\Pi}^\top(\boldsymbol{y} - \boldsymbol{Z}\widehat{\boldsymbol{\varepsilon}}) \tag{2.63}$$

and

$$
\begin{aligned}
\widehat{\boldsymbol{\varepsilon}} &= \arg\min_{\boldsymbol{\varepsilon}\in\mathbb{R}^{n-2}} \|\boldsymbol{Q}_{\Pi}\boldsymbol{y} - \boldsymbol{Q}_{\Pi}\boldsymbol{Z}\boldsymbol{\varepsilon}\|^2 + \lambda\|\boldsymbol{\varepsilon}\|^2 \\
&= (\boldsymbol{Z}^{\top}\boldsymbol{Q}_{\Pi}\boldsymbol{Z} + \lambda\boldsymbol{I}_{n-2})^{-1}\boldsymbol{Z}^{\top}\boldsymbol{Q}_{\Pi}\boldsymbol{y}, \qquad (2.64)
\end{aligned}
$$

which shows that we may obtain (penalized) regressions relating to the cubic smoothing spline even if we use the other right-inverse matrices of $\boldsymbol{D}$ such that $\boldsymbol{Z} \in \Gamma_D\backslash\{\boldsymbol{D}_{\mathrm{r}}^{-1}\}$. Nevertheless, as illustrated here, they are more complex than those shown in Tables 1–2.

2.7 Concluding remarks

In this chapter, we provided a comprehensive list of penalized least squares regressions relating to the cubic smoothing spline, and then revealed a principle of duality in them. This is the main contribution of this study. Such penalized regressions are tabulated in Tables 1–2 and the principle of duality revealed is stated in Proposition 2.2. In addition, we also provided a number of results derived from them, most of which are also tabulated in Tables 1–2 and some of which are illustrated in Figures 2.1–2.4.

2.8 Appendix

2.8.1 Some remarks on a special case such that $\boldsymbol{x} = [1, \dots, n]^\top$

(i) If $\boldsymbol{x} = [1, \dots, n]^\top$, then $\boldsymbol{C} = \boldsymbol{D}_{(2)}\boldsymbol{D}_{(1)} \in \mathbb{R}^{(n-2)\times n}$, which is a Toeplitz matrix whose first (resp. last) row is $[1, -2, 1, 0, \dots, 0]$ (resp. $[0, \dots, 0, 1, -2, 1]$). (ii) If $\boldsymbol{x} = [1, \dots, n]^\top$, then $(\boldsymbol{I}_n + \lambda \boldsymbol{C}^\top \boldsymbol{R}^{-1}\boldsymbol{C})^{-1}$ is bisymmetric (i.e., symmetric centrosymmetric), which may be proved as in Yamada (2020a). (iii) If $\boldsymbol{x} = [1, \dots, n]^\top$, then $\boldsymbol{R}$ in (2.8) is not only a symmetric tridiagonal matrix but also a Toeplitz matrix. In the case, we have

$$\omega_k = \frac{2}{3} + \frac{1}{3}\cos\left(\frac{k\pi}{n-1}\right), \quad k = 1, \dots, n-2, \tag{2.65}$$

and thus ω_{n-2}, which is the smallest eigenvalue of $\boldsymbol{R}$, satisfies the following inequality [see, e.g., Pesaran (1973)]:

$$\omega_{n-2} = \frac{2}{3} + \frac{1}{3}\cos\left(\frac{n-2}{n-1}\pi\right) > \frac{1}{3}. \tag{2.66}$$

(iv) If $\boldsymbol{x} = [1, \dots, n]^\top$ and $\boldsymbol{R} = \boldsymbol{I}_{n-2}$ in (2.2)–(2.3), then (2.2)–(2.3) reduce to

$$\begin{aligned}\widehat{\boldsymbol{f}} &= \arg \min_{\boldsymbol{f}\in\mathbb{R}^n} \|\boldsymbol{y} - \boldsymbol{f}\|^2 + \|\boldsymbol{D}_{(2)}\boldsymbol{D}_{(1)}\boldsymbol{f}\|^2 \\ &= \left\{\boldsymbol{I}_n + \lambda(\boldsymbol{D}_{(2)}\boldsymbol{D}_{(1)})^\top(\boldsymbol{D}_{(2)}\boldsymbol{D}_{(1)})\right\}^{-1}\boldsymbol{y}. \end{aligned} \tag{2.67}$$

It is a type of the Whittaker–Henderson (WH) method of graduation, which was developed by Bohlmann (1899), Whittaker (1923) and others. See Weinert (2007) for a historical review of the WH method of graduation. (2.67) is also referred to as the Hodrick–Prescott (HP) (1997) filtering in econometrics. For more details about the HP filtering, see, e.g., Schlicht (2005), Kim et al. (2009), Paige and Trindade (2010), and Yamada (2015, 2018ab, 2020b).

2.8.2 User-defined functions

2.8.2.1 A Matlab/GNU Octave function to make $\boldsymbol{C}$ in (2.7)

```
function C=makeCmat(x)
```

```
    n=length(x); D1=diff(eye(n)); D2=diff(eye(n-1));
    delta=diff(x); invDelta=diag(1./delta);
    C=D2*invDelta*D1;
end
```

2.8.2.2 A Matlab/GNU Octave function to make R in (2.8)

```
function R=makeRmat(x)
    n=length(x); delta=diff(x);
    R0=diag(delta(1:n-2)+delta(2:n-1))/3;
    R1=diag(delta(2:n-2),1)/6;
    R=R1'+R0+R1;
end
```

2.8.2.3 A Matlab/GNU Octave function to make D in (2.9)

```
function D=makeDmat(x)
    C=makeCmat(x); R=makeRmat(x); [P,L]=eig(R);
    invsqrtR=P*diag(sqrt(diag(L)))*P';
    D=invsqrtR*C;
end
```

2.8.2.4 A R function to make C in (2.7)

```
makeCmat <- function(x) {
# Note: x is an n x 1 matrix (not a vector).
    n <- length(x)
    D1 <- diff(diag(n)); D2 <- diff(diag(n-1))
    delta <- diff(x); invDelta <- diag(1/delta[1:(n-1),1])
    C <- D2%*%invDelta%*%D1
    return(C)
}
```

2.8.2.5 A R function to make R in (2.8)

```
makeRmat <- function(x) {
# Note: x is an n x 1 matrix (not a vector).
```

```
  n <- length(x); delta <- diff(x)
  R0 <- diag((delta[1:(n-2),1]+delta[2:(n-1),1])/3)
  R1 <- diag(0,n-2)
  R1[row(R1)==col(R1)-1] <- delta[2:(n-2),1]/6
  R <- t(R1)+R0+R1
  return(R)
}
```

2.8.2.6 A R function to make D in (2.9)

```
makeDmat <- function(x) {
# Note: x is an n x 1 matrix (not a vector).
  n <- length(x); C <- makeCmat(x); R <- makeRmat(x)
  z <- eigen(R); P <- z$vectors
  invsqrtR <- P%*%diag(sqrt(z$values))%*%t(P)
  D <- invsqrtR%*%C
  return(D)
}
```

Table 1: Most of the main results (I)

	Regressions relating to the cubic smoothing spline	Average	$\lambda \to \infty$	$\lambda \to 0$	$\boldsymbol{y} \in \mathcal{S}(\mathcal{P})$	Sum	$\perp$
(P1)	$\widehat{\boldsymbol{f}}(=\widehat{\boldsymbol{\tau}}+\widehat{\boldsymbol{c}}) = \arg\min_{\boldsymbol{f}\in\mathbb{R}^n} \lVert\boldsymbol{y}-\boldsymbol{f}\rVert^2+\lambda\lVert\mathcal{A}\boldsymbol{f}\rVert^2 = \left(\boldsymbol{I}_n+\lambda\mathcal{A}^\top\mathcal{A}\right)^{-1}\boldsymbol{y}$	$\bar{y}$	$\widehat{\boldsymbol{\tau}}$	$\boldsymbol{y}$	$\boldsymbol{y}$	1	
(D1)	$\widehat{\boldsymbol{g}}(=\widehat{\boldsymbol{\tau}}+\widehat{\boldsymbol{u}}) = \arg\min_{\boldsymbol{g}\in\mathbb{R}^n} \lVert\boldsymbol{y}-\boldsymbol{g}\rVert^2+\lambda^{-1}\lVert\mathcal{A}_{\mathrm{r}}^{-1\top}\boldsymbol{g}\rVert^2 = \left(\boldsymbol{I}_n+\lambda^{-1}\mathcal{A}_{\mathrm{r}}^{-1}\mathcal{A}_{\mathrm{r}}^{-1\top}\right)^{-1}\boldsymbol{y}$	$\bar{y}$	$\boldsymbol{y}$	$\widehat{\boldsymbol{\tau}}$	$\boldsymbol{y}$	1	
(P2)	$\widehat{\boldsymbol{u}} = \mathcal{A}^\top\widehat{\boldsymbol{\eta}}$, where $\widehat{\boldsymbol{\eta}} = \arg\min_{\boldsymbol{\eta}\in\mathbb{R}^{n-2}} \lVert\boldsymbol{y}-\mathcal{A}^\top\boldsymbol{\eta}\rVert^2+\lambda^{-1}\lVert\boldsymbol{\eta}\rVert^2 = (\mathcal{A}\mathcal{A}^\top+\lambda^{-1}\boldsymbol{I}_{n-2})^{-1}\mathcal{A}\boldsymbol{y}$	0	$\boldsymbol{y}-\widehat{\boldsymbol{\tau}}$	$\boldsymbol{0}$	$\boldsymbol{0}$	0	$\circ$
(D2)	$\widehat{\boldsymbol{c}} = \mathcal{A}_{\mathrm{r}}^{-1}\widehat{\boldsymbol{\gamma}}$, where $\widehat{\boldsymbol{\gamma}} = \arg\min_{\boldsymbol{\gamma}\in\mathbb{R}^{n-2}} \lVert\boldsymbol{y}-\mathcal{A}_{\mathrm{r}}^{-1}\boldsymbol{\gamma}\rVert^2+\lambda\lVert\boldsymbol{\gamma}\rVert^2 = (\mathcal{A}_{\mathrm{r}}^{-1\top}\mathcal{A}_{\mathrm{r}}^{-1}+\lambda\boldsymbol{I}_{n-2})^{-1}\mathcal{A}_{\mathrm{r}}^{-1\top}\boldsymbol{y}$	0	$\boldsymbol{0}$	$\boldsymbol{y}-\widehat{\boldsymbol{\tau}}$	$\boldsymbol{0}$	0	$\circ$
(P3)	$\widehat{\boldsymbol{c}} = \arg\min_{\boldsymbol{c}\in\mathbb{R}^n} \lVert(\boldsymbol{y}-\widehat{\boldsymbol{\tau}})-\boldsymbol{c}\rVert^2+\lambda\lVert\mathcal{A}\boldsymbol{c}\rVert^2 = \left(\boldsymbol{I}_n+\lambda\mathcal{A}^\top\mathcal{A}\right)^{-1}(\boldsymbol{y}-\widehat{\boldsymbol{\tau}})$	0	$\boldsymbol{0}$	$\boldsymbol{y}-\widehat{\boldsymbol{\tau}}$	$\boldsymbol{0}$	0	$\circ$
(D3)	$\widehat{\boldsymbol{u}} = \arg\min_{\boldsymbol{u}\in\mathbb{R}^n} \lVert(\boldsymbol{y}-\widehat{\boldsymbol{\tau}})-\boldsymbol{u}\rVert^2+\lambda^{-1}\lVert\mathcal{A}_{\mathrm{r}}^{-1\top}\boldsymbol{u}\rVert^2 = \left(\boldsymbol{I}_n+\lambda^{-1}\mathcal{A}_{\mathrm{r}}^{-1}\mathcal{A}_{\mathrm{r}}^{-1\top}\right)^{-1}(\boldsymbol{y}-\widehat{\boldsymbol{\tau}})$	0	$\boldsymbol{y}-\widehat{\boldsymbol{\tau}}$	$\boldsymbol{0}$	$\boldsymbol{0}$	0	$\circ$
(P4)	$\widehat{\boldsymbol{h}}(=\widehat{\boldsymbol{c}}+\widehat{\boldsymbol{u}}) = \mathcal{D}^\top\widehat{\boldsymbol{\alpha}}$, where $\widehat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}\in\mathbb{R}^{n-2}} \lVert\boldsymbol{y}-\mathcal{D}^\top\boldsymbol{\alpha}\rVert^2 = (\mathcal{D}\mathcal{D}^\top)^{-1}\mathcal{D}\boldsymbol{y}$	0	$\boldsymbol{y}-\widehat{\boldsymbol{\tau}}$	$\boldsymbol{y}-\widehat{\boldsymbol{\tau}}$	$\boldsymbol{0}$	0	$\circ$
(D4)	$\widehat{\boldsymbol{h}}(=\widehat{\boldsymbol{c}}+\widehat{\boldsymbol{u}}) = \mathcal{D}_{\mathrm{r}}^{-1}\widehat{\boldsymbol{\rho}}$, where $\widehat{\boldsymbol{\rho}} = \arg\min_{\boldsymbol{\rho}\in\mathbb{R}^{n-2}} \lVert\boldsymbol{y}-\mathcal{D}_{\mathrm{r}}^{-1}\boldsymbol{\rho}\rVert^2 = (\mathcal{D}_{\mathrm{r}}^{-1\top}\mathcal{D}_{\mathrm{r}}^{-1})^{-1}\mathcal{D}_{\mathrm{r}}^{-1\top}\boldsymbol{y}$	0	$\boldsymbol{y}-\widehat{\boldsymbol{\tau}}$	$\boldsymbol{y}-\widehat{\boldsymbol{\tau}}$	$\boldsymbol{0}$	0	$\circ$
	$\widehat{\boldsymbol{\tau}} = \mathcal{P}\widehat{\boldsymbol{\beta}}$, where $\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^2} \lVert\boldsymbol{y}-\mathcal{P}\boldsymbol{\beta}\rVert^2 = (\mathcal{P}^\top\mathcal{P})^{-1}\mathcal{P}^\top\boldsymbol{y}$	$\bar{y}$	$\widehat{\boldsymbol{\tau}}$	$\widehat{\boldsymbol{\tau}}$	$\boldsymbol{y}$	1	

- $\boldsymbol{y} = \widehat{\boldsymbol{\tau}} + \widehat{\boldsymbol{c}} + \widehat{\boldsymbol{u}}$.
- $\mathcal{A} = \boldsymbol{D}, \boldsymbol{F}$.
- $\mathcal{D} = \boldsymbol{C}, \boldsymbol{D}, \boldsymbol{E}, \boldsymbol{F}$.
- $\boldsymbol{M}_{\mathrm{r}}^{-1} = \boldsymbol{M}^\top(\boldsymbol{M}\boldsymbol{M}^\top)^{-1}$ for $\boldsymbol{M} = \mathcal{A}, \mathcal{D}$.
- $\mathcal{P} = \boldsymbol{\Pi}, \boldsymbol{T}$.
- $\lambda > 0$ is a smoothing/tuning parameter.
- $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$.
- $\mathcal{S}(\mathcal{P})$ denotes the column space of $\mathcal{P}$.
- 'Sum' denotes the sum of the entries in each row of the hat matrices.
- $\circ$ indicates that the corresponding component belongs to the orthogonal complement of $\mathcal{S}(\mathcal{P})$.

Table 2: Most of the main results (II)

	Regressions relating to the cubic smoothing spline	Average	$\lambda \to \infty$	$\lambda \to 0$	$\boldsymbol{y} \in \mathcal{S}(\mathcal{P})$	Sum	$\perp$
(P5)	$\widehat{\boldsymbol{f}}(=\widehat{\boldsymbol{\tau}}+\widehat{\boldsymbol{c}}) = \arg\min_{\boldsymbol{f}\in\mathbb{R}^n} \\|\boldsymbol{y}-\boldsymbol{f}\\|^2+\lambda\boldsymbol{f}^\top\mathcal{B}^\top\mathcal{Q}^{-1}\mathcal{B}\boldsymbol{f} = \left(\boldsymbol{I}_n+\lambda\mathcal{B}^\top\mathcal{Q}^{-1}\mathcal{B}\right)^{-1}\boldsymbol{y}$	$\bar{y}$	$\widehat{\boldsymbol{\tau}}$	$\boldsymbol{y}$	$\boldsymbol{y}$	1	
(D5)	$\widehat{\boldsymbol{g}}(=\widehat{\boldsymbol{\tau}}+\widehat{\boldsymbol{u}}) = \arg\min_{\boldsymbol{g}\in\mathbb{R}^n} \\|\boldsymbol{y}-\boldsymbol{g}\\|^2+\lambda^{-1}\boldsymbol{g}^\top\mathcal{B}_{\mathrm{r}}^{-1}\mathcal{Q}\mathcal{B}_{\mathrm{r}}^{-1\top}\boldsymbol{g} = \left(\boldsymbol{I}_n+\lambda^{-1}\mathcal{B}_{\mathrm{r}}^{-1}\mathcal{Q}\mathcal{B}_{\mathrm{r}}^{-1\top}\right)^{-1}\boldsymbol{y}$	$\bar{y}$	$\boldsymbol{y}$	$\widehat{\boldsymbol{\tau}}$	$\boldsymbol{y}$	1	
(P6)	$\widehat{\boldsymbol{u}} = \mathcal{B}^\top\widehat{\boldsymbol{v}}$, where $\widehat{\boldsymbol{v}} = \arg\min_{\boldsymbol{v}\in\mathbb{R}^{n-2}} \\|\boldsymbol{y}-\mathcal{B}^\top\boldsymbol{v}\\|^2+\lambda^{-1}\boldsymbol{v}^\top\mathcal{Q}\boldsymbol{v} = (\mathcal{B}\mathcal{B}^\top+\lambda^{-1}\mathcal{Q})^{-1}\mathcal{B}\boldsymbol{y}$	0	$\boldsymbol{y}-\widehat{\boldsymbol{\tau}}$	$\boldsymbol{0}$	$\boldsymbol{0}$	0	$\circ$
(D6)	$\widehat{\boldsymbol{c}} = \mathcal{B}_{\mathrm{r}}^{-1}\widehat{\boldsymbol{\kappa}}$, where $\widehat{\boldsymbol{\kappa}} = \arg\min_{\boldsymbol{\kappa}\in\mathbb{R}^{n-2}} \\|\boldsymbol{y}-\mathcal{B}_{\mathrm{r}}^{-1}\boldsymbol{\kappa}\\|^2+\lambda\boldsymbol{\kappa}^\top\mathcal{Q}^{-1}\boldsymbol{\kappa} = (\mathcal{B}_{\mathrm{r}}^{-1\top}\mathcal{B}_{\mathrm{r}}^{-1}+\lambda\mathcal{Q}^{-1})^{-1}\mathcal{B}_{\mathrm{r}}^{-1\top}\boldsymbol{y}$	0	$\boldsymbol{0}$	$\boldsymbol{y}-\widehat{\boldsymbol{\tau}}$	$\boldsymbol{0}$	0	$\circ$
(P7)	$\widehat{\boldsymbol{c}} = \arg\min_{\boldsymbol{c}\in\mathbb{R}^n} \\|(\boldsymbol{y}-\widehat{\boldsymbol{\tau}})-\boldsymbol{c}\\|^2+\lambda\boldsymbol{c}^\top\mathcal{B}^\top\mathcal{Q}^{-1}\mathcal{B}\boldsymbol{c} = \left(\boldsymbol{I}_n+\lambda\mathcal{B}^\top\mathcal{Q}^{-1}\mathcal{B}\right)^{-1}(\boldsymbol{y}-\widehat{\boldsymbol{\tau}})$	0	$\boldsymbol{0}$	$\boldsymbol{y}-\widehat{\boldsymbol{\tau}}$	$\boldsymbol{0}$	0	$\circ$
(D7)	$\widehat{\boldsymbol{u}} = \arg\min_{\boldsymbol{u}\in\mathbb{R}^n} \\|(\boldsymbol{y}-\widehat{\boldsymbol{\tau}})-\boldsymbol{u}\\|^2+\lambda^{-1}\boldsymbol{u}^\top\mathcal{B}_{\mathrm{r}}^{-1}\mathcal{Q}\mathcal{B}_{\mathrm{r}}^{-1\top}\boldsymbol{u} = \left(\boldsymbol{I}_n+\lambda^{-1}\mathcal{B}_{\mathrm{r}}^{-1}\mathcal{Q}\mathcal{B}_{\mathrm{r}}^{-1\top}\right)^{-1}(\boldsymbol{y}-\widehat{\boldsymbol{\tau}})$	0	$\boldsymbol{y}-\widehat{\boldsymbol{\tau}}$	$\boldsymbol{0}$	$\boldsymbol{0}$	0	$\circ$

- $\boldsymbol{y} = \widehat{\boldsymbol{\tau}}+\widehat{\boldsymbol{c}}+\widehat{\boldsymbol{u}}$.
- $(\mathcal{B},\mathcal{Q}) = (\boldsymbol{C},\boldsymbol{R}),(\boldsymbol{E},\boldsymbol{S}^{-1})$.
- $\mathcal{B}_{\mathrm{r}}^{-1} = \mathcal{B}^\top(\mathcal{B}\mathcal{B}^\top)^{-1}$.
- $\lambda > 0$ is a smoothing/tuning parameter.
- $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$.
- $\mathcal{S}(\mathcal{P})$ denotes the column space of $\mathcal{P}$, where $\mathcal{P} = \boldsymbol{\Pi},\boldsymbol{T}$.
- 'Sum' denotes the sum of the entries in each row of the hat matrices.
- $\circ$ indicates that the corresponding component belongs to the orthogonal complement of $\mathcal{S}(\mathcal{P})$.

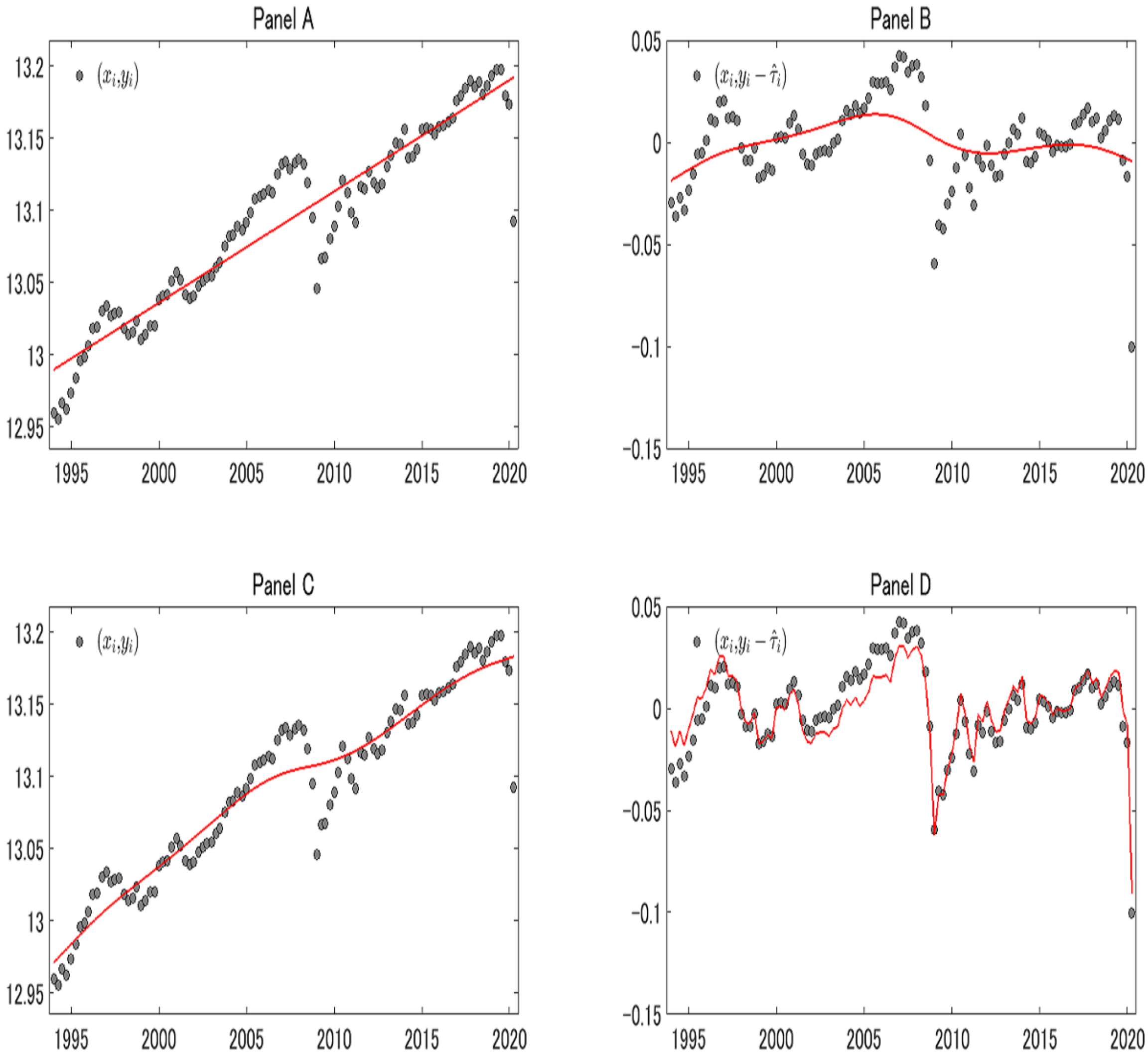

Figure 2.1: Panel A shows a scatter plot of the log of seasonally adjusted Japanese real gross domestic product (GDP) over the sample period 1994:Q1 to 2020:Q2. The solid line in the panel plots $(x_i, \widehat{\tau}_i)$ for $i = 1, \ldots, n$, where $\widehat{\boldsymbol{\tau}} = [\widehat{\tau}_1, \ldots, \widehat{\tau}_n]^\top$ in (2.4) and $n = 106$. Panel B depicts a scatter plot of $(x_i, y_i - \widehat{\tau}_i)$ for $i = 1, \ldots, n$. The solid line in the panel plots $(x_i, \widehat{c}_i)$ for $i = 1, \ldots, n$, where $\widehat{\boldsymbol{c}} = [\widehat{c}_1, \ldots, \widehat{c}_n]^\top$ is calculated by (2.18) with $\lambda = 10^3$. The solid line in Panel C denotes $(x_i, \widehat{f}_i)$, where $\widehat{\boldsymbol{f}} = [\widehat{f}_1, \ldots, \widehat{f}_n]^\top$ is calculated by (2.14) with $\lambda = 10^3$. Panel D illustrates a scatter plot of $(x_i, y_i - \widehat{\tau}_i)$ for $i = 1, \ldots, n$. The solid line in the panel plots $(x_i, \widehat{u}_i)$ for $i = 1, \ldots, n$, where $\widehat{\boldsymbol{u}} = [\widehat{u}_1, \ldots, \widehat{u}_n]^\top$ is calculated by (2.27) with $\lambda = 1600$.

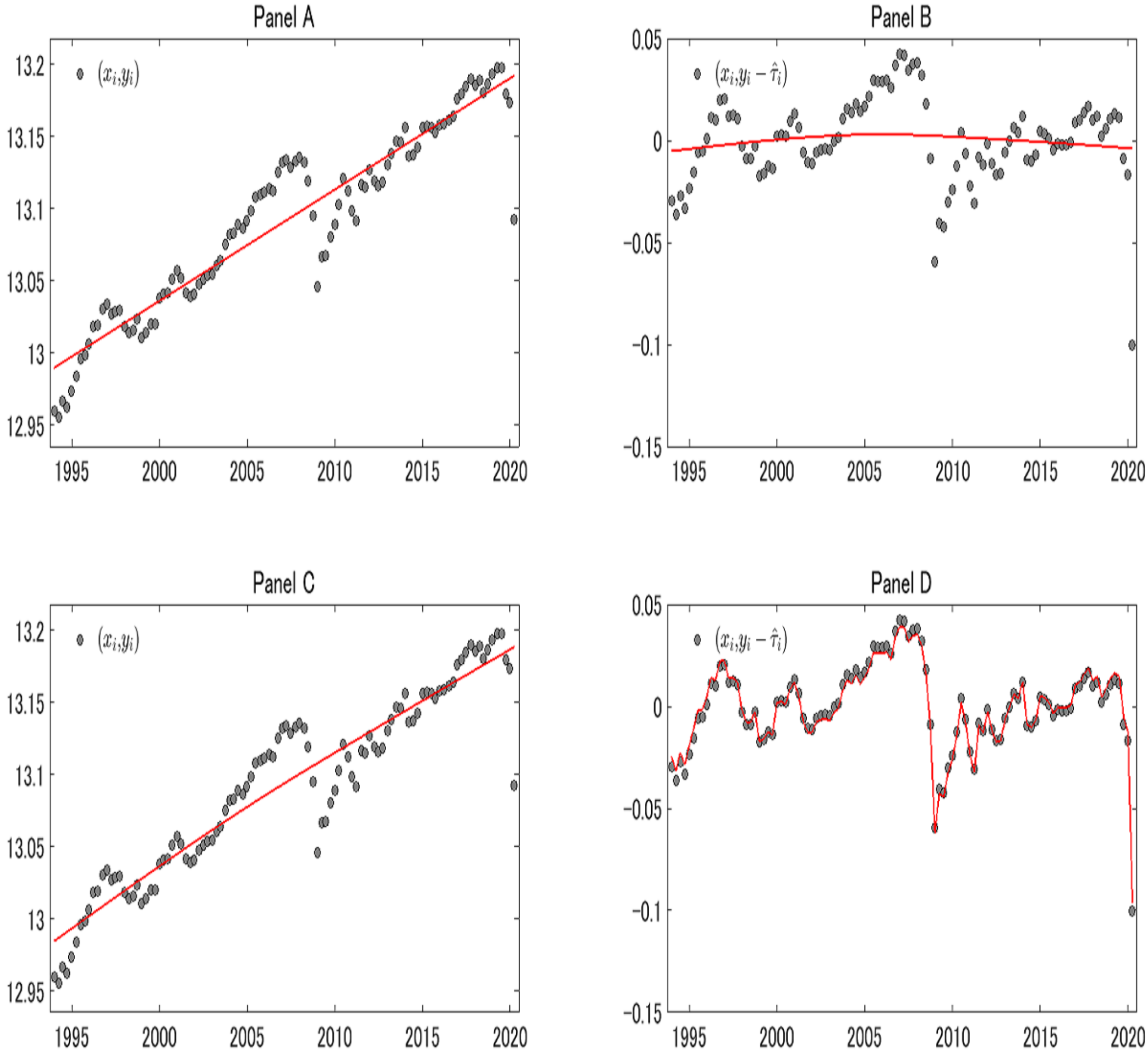

Figure 2.2: This figure corresponds to the case where $\lambda = 10^5$. For the other explanations, see Figure 2.1.

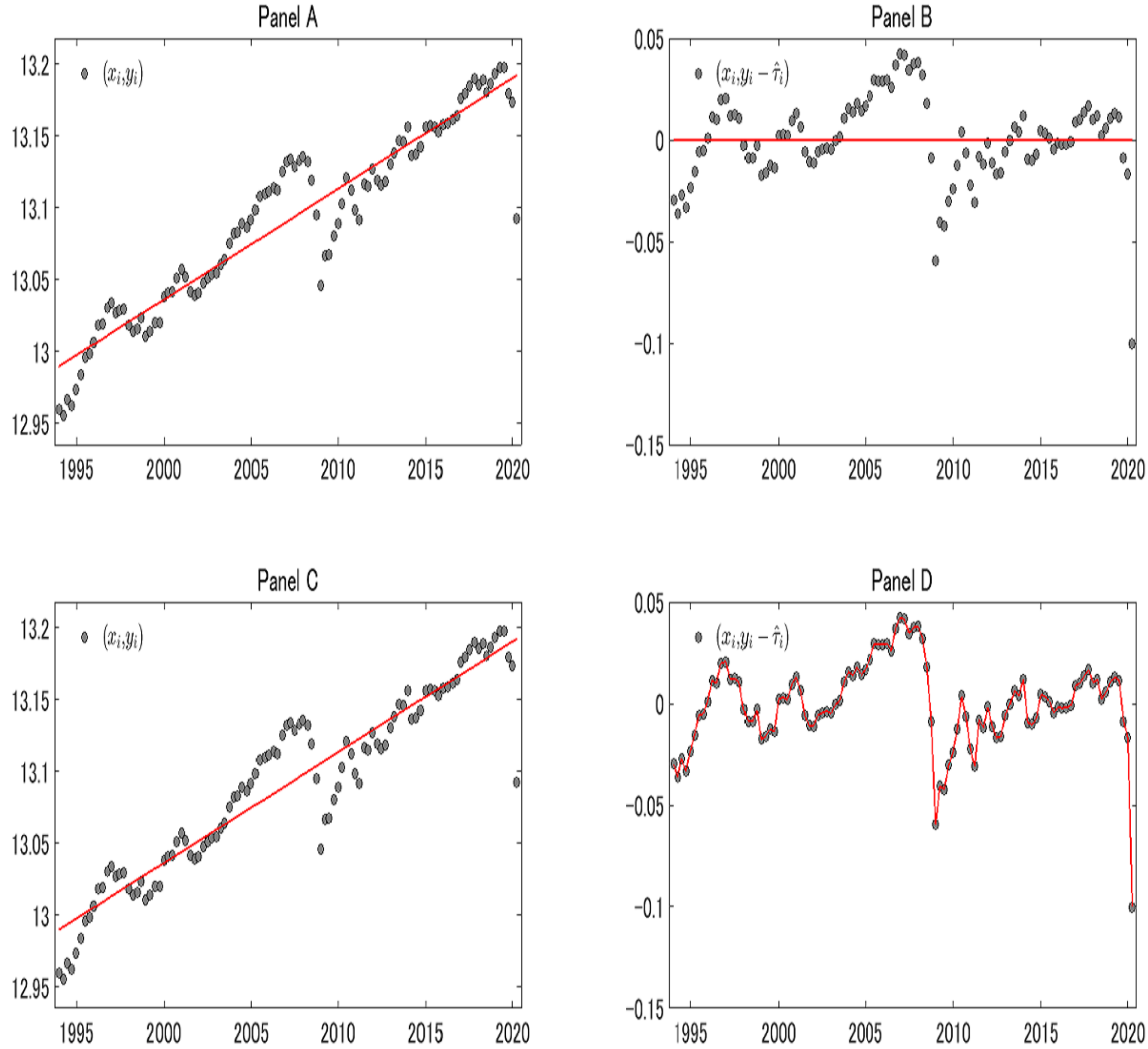

Figure 2.3: This figure corresponds to the case where $\lambda = 10^{10}$. For the other explanations, see Figure 2.1.

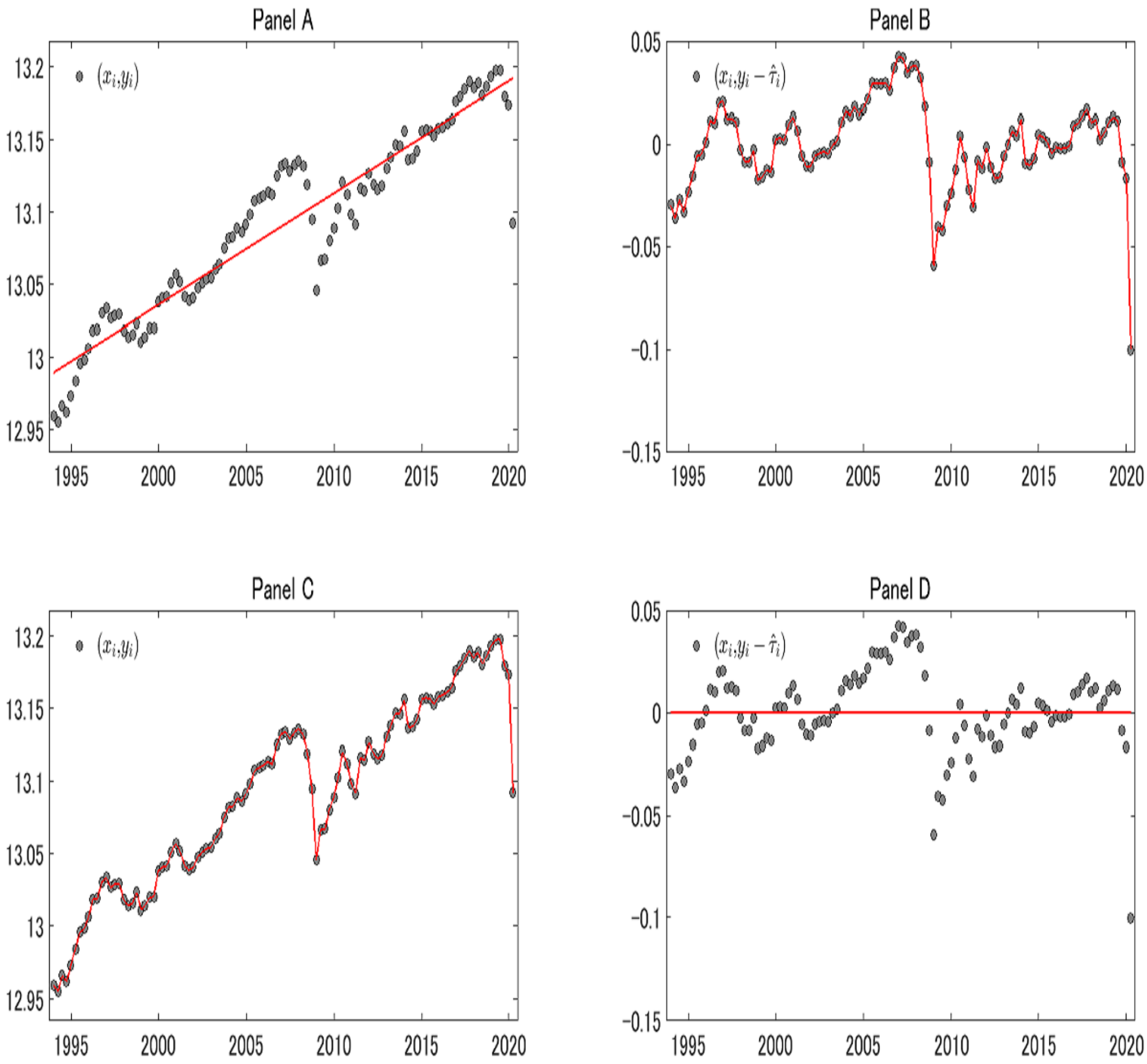

Figure 2.4: This figure corresponds to the case where $\lambda = 10^{-10}$. For the other explanations, see Figure 2.1.

2.9 References

1. Bohlmann, G., 1899, Ein ausgleichungsproblem, *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse*, 260–271.

2. Du, R. and H. Yamada, 2020, Principle of duality in cubic smoothing spline, *Mathematics*, 8, 10, 1839.

3. Green, P. J. and B. W. Silverman, 1994, *Nonparametric Regression and Generalized Linear Models: A roughness Penalty Approach*, Chapman and Hall/CRC, Boca Raton.

4. Hodrick, R. J. and E. C. Prescott, 1997, Postwar U.S. business cycles: An empirical investigation, *Journal of Money, Credit and Banking*, 29, 1, 1–16.

5. Hoerl, A. E. and R. W. Kennard, 1970, Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics*, 12, 1, 55–67.

6. Høyer, J. L. and I. Karagali, 2016, Sea surface temperature climate data record for the North Sea and Baltic Sea, *Journal of Climate*, 29, 7, 2529–2541.

7. Kim, S., K. Koh, S. Boyd, and D. Gorinevsky, 2009, ℓ_1 trend filtering, *SIAM Review*, 51, 2, 339–360.

8. Paige, R. L. and A. A. Trindade, 2010, The Hodrick–Prescott filter: A special case of penalized spline smoothing, *Electronic Journal of Statistics*, 4, 856–874.

9. Pesaran, M. H., 1973, Exact maximum likelihood estimation of a regression equation with a first-order moving-average error, *Review of Economic Studies*, 40, 4, 529–535.

10. Reinsch, C., 1967, Smoothing by spline functions, *Numerische Mathematik*, 10, 3, 177–183.

11. Schlicht, E., 2005, Estimating the smoothing parameter in the so-called Hodrick–Prescott filter, *Journal of the Japan Statistical Society*, 35, 1, 99–119.

12. Schoenberg, I. J., 1964, Spline functions and the problem of graduation, *Proceedings of the National Academy of Sciences*, 52, 947–950.

13. Verbyla, A. P., B. R. Cullis, M. G. Kenward, and S. J. Welham, 1999, The analysis of designed experiments and longitudinal data by using smoothing splines, *Journal of the Royal Statistical Society C*, 48, 3, 269–311.

14. Weinert, H. L., 2007, Efficient computation for Whittaker–Henderson smoothing, *Computational Statistics and Data Analysis*, 52, 2, 959–974.

15. Whittaker, E. T., 1923, On a new method of graduation, *Proceedings of the Edinburgh Mathematical Society*, 41, 63–75.

16. Wood, S. N., 2017, *Generalized Additive Models: An Introduction with R*, second edition, Chapman & Hall/CRC, Boca Raton.

17. Yamada, H., 2015, Ridge regression representations of the generalized Hodrick–Prescott filter, *Journal of the Japan Statistical Society*, 45, 2, 121–128.

18. Yamada, H., 2017, The Frisch–Waugh–Lovell theorem for the lasso and the ridge regression, *Communications in Statistics – Theory and Methods*, 46, 21, 10897–10902.

19. Yamada, H., 2018a, Why does the trend extracted by the Hodrick–Prescott filtering seem to be more plausible than the linear trend?, *Applied Economics Letters*, 25, 2, 102–105.

20. Yamada, H., 2018b, Several least squares problems related to the Hodrick–Prescott filtering, *Communications in Statistics – Theory and Methods*, 47, 5, 1022–1027.

21. Yamada, H., 2020a, A note on Whittaker–Henderson graduation: Bisymmetry of the smoother matrix, *Communications in Statistics – Theory and Methods*, 49, 7, 1629–1634.

22. Yamada, H., 2020b, A smoothing method that looks like the Hodrick–Prescott filter, *Econometric Theory*, 36, 5, 961–981.

Chapter 3

A Modification of the Whittaker–Henderson Method of Graduation

This chapter is based on a previously published article: Yamada and Du (2019).

3.1 Introduction

The squared ℓ_2-norm penalized least squares problem defined as

$$(\widetilde{x}_1, \ldots, \widetilde{x}_T) = \underset{x_1,\ldots,x_T \in \mathbb{R}}{\arg\min} \sum_{t=1}^{T} (y_t - x_t)^2 + \lambda \sum_{t=3}^{T} (\Delta^2 x_t)^2, \tag{3.1}$$

where $y_1, \ldots, y_T$ are observed time series data, has been referred to as the Hodrick–Prescott (HP) filter in econometrics since its use by Hodrick and Prescott (1997). Here, $\lambda > 0$ and Δ denotes the backward difference operator such that $\Delta x_t = x_t - x_{t-1}$. It is applied to decompose y_t for $t = 1, \ldots, T$ into $\widetilde{x}_t$ (the trend) and $\widetilde{c}_t = y_t - \widetilde{x}_t$.

Yamada (2017) recently introduced the following modification:

$$(\widehat{x}_1, \ldots, \widehat{x}_T, \widehat{x}_{T+1}, \ldots, \widehat{x}_{T+h}) = \underset{x_1,\ldots,x_T,x_{T+1},\ldots,x_{T+h} \in \mathbb{R}}{\arg\min} \sum_{t=1}^{T} (y_t - x_t)^2 + \lambda \sum_{t=3}^{T+h} (\Delta^2 x_t)^2, \tag{3.2}$$

and showed that

$$
\begin{cases}
\widehat{x}_t = \widetilde{x}_t, & (t = 1, \ldots, T), \\
\widehat{x}_{T+j} = \widetilde{x}_T + j(\widetilde{x}_T - \widetilde{x}_{T-1}), & (j = 1, \ldots, h).
\end{cases}
\tag{3.3}
$$

Thus the above filter, (3.2), provides not only identical trend estimates to those of the HP filter, but also extrapolations of the trend beyond the sample limit (taken as $t = T$), and is therefore of practical use. In addition, Yamada (2017) showed that

$$
\lim_{\lambda \to \infty} \widehat{x}_t = \widehat{\alpha}_0 + \widehat{\alpha}_1 t, \quad (t = 1, \ldots, T, T+1, \ldots, T+h),
\tag{3.4}
$$

where $(\widehat{\alpha}_0, \widehat{\alpha}_1) = \arg\min_{\alpha_0, \alpha_1 \in \mathbb{R}} \sum_{t=1}^{T} (y_t - \alpha_0 - \alpha_1 t)^2$.

The HP filter in (3.1) is a special case of the Whittaker–Henderson (WH) method of graduation:

$$
(\widetilde{z}_1, \ldots, \widetilde{z}_T) = \underset{z_1, \ldots, z_T \in \mathbb{R}}{\arg\min} \sum_{t=1}^{T} (y_t - z_t)^2 + \lambda \sum_{t=p+1}^{T} (\Delta^p z_t)^2,
\tag{3.5}
$$

which was developed by Whittaker (1923) and others. For historical survey, see Weinert (2007), Phillips (2010) and Nocon and Scott (2012).

Corresponding to the modification from (3.1) to (3.2), (3.5) may be generalized as follows:

$$
(\widehat{z}_1, \ldots, \widehat{z}_T, \widehat{z}_{T+1}, \ldots, \widehat{z}_{T+h}) = \underset{z_1, \ldots, z_T, z_{T+1}, \ldots, z_{T+h} \in \mathbb{R}}{\arg\min} \sum_{t=1}^{T} (y_t - z_t)^2 + \lambda \sum_{t=p+1}^{T+h} (\Delta^p z_t)^2.
\tag{3.6}
$$

In this chapter, after presenting the closed-form solution of the modified WH graduation, (3.6), we prove generalizations of (3.3) and (3.4). In addition, we introduce two other penalized least squares problems and show that they lead to the same results as those of the modified WH graduation, (3.6).

Notations $\boldsymbol{I}_T \in \mathbb{R}^{T \times T}$ is an identity matrix, $\mathbf{0}_{m,T} \in \mathbb{R}^{m \times T}$ is a zero matrix, $\boldsymbol{S}_T = [\boldsymbol{I}_T, \mathbf{0}_{T,h}] \in \mathbb{R}^{T \times (T+h)}$, $\boldsymbol{\Pi}_T \in \mathbb{R}^{T \times p}$ is a matrix such that its t-th row is $[1, t, \ldots, t^{p-1}]$ for $t = 1, \ldots, n$, and $\boldsymbol{D}_T \in \mathbb{R}^{(n-p) \times T}$ is a p-th order difference matrix such that $\boldsymbol{D}_T \boldsymbol{\eta} = [\Delta^p \eta_{p+1}, \ldots, \Delta^p \eta_n]^\top$ for an n-dimensional column vector $\boldsymbol{\eta} = [\eta_1, \ldots, \eta_n]^\top$.

3.2 A modification of the WH graduation

Letting $\boldsymbol{y} = [y_1, \dots, y_T]^\top$, $\boldsymbol{z} = [z_1, \dots, z_T]^\top$, and $\widetilde{\boldsymbol{z}} = [\widetilde{z}_1, \dots, \widetilde{z}_T]^\top$, The objective function of (3.5) may be represented in matrix notation as $\|\boldsymbol{y} - \boldsymbol{z}\|^2 + \lambda\|\boldsymbol{D}_T \boldsymbol{z}\|^2$, and $\widetilde{\boldsymbol{z}}$ may be expressed explicitly as

$$\widetilde{\boldsymbol{z}} = (\boldsymbol{I}_T + \lambda \boldsymbol{D}_T^\top \boldsymbol{D}_T)^{-1} \boldsymbol{y}. \tag{3.7}$$

Letting $\boldsymbol{v} = [z_1, \dots, z_T, z_{T+1}, \dots, z_{T+h}]^\top$ and $\widehat{\boldsymbol{z}} = [\widehat{z}_1, \dots, \widehat{z}_T, \widehat{z}_{T+1}, \dots, \widehat{z}_{T+h}]^\top$, the modified WH graduation is represented in matrix notation as $\widehat{\boldsymbol{z}} = \arg\min_{\boldsymbol{v} \in \mathbb{R}^{T+h}} \|\boldsymbol{y} - \boldsymbol{S}_T \boldsymbol{v}\|^2 + \lambda \|\boldsymbol{D}_{T+h} \boldsymbol{v}\|^2$. We obtain the following closed-form solution of the modified WH graduation:[1]

$$\widehat{\boldsymbol{z}} = (\boldsymbol{S}_T^\top \boldsymbol{S}_T + \lambda \boldsymbol{D}_{T+h}^\top \boldsymbol{D}_{T+h})^{-1} \boldsymbol{S}_T^\top \boldsymbol{y}. \tag{3.8}$$

We note that $\boldsymbol{D}_{T+h}$ is a $(p+1)$-diagonal Toeplitz matrix such that

$$\boldsymbol{D}_{T+h} = \begin{bmatrix} a_0 & \cdots & a_p & 0 & \cdots & 0 \\ 0 & \ddots & & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & & \ddots & 0 \\ 0 & \cdots & 0 & a_0 & \cdots & a_p \end{bmatrix},$$

where $a_k = (-1)^{p-k}\binom{p}{k}$ for $k = 0, \dots, p$, and thus may be expressed as

$$\boldsymbol{D}_{T+h} = \begin{bmatrix} \boldsymbol{D}_T & \boldsymbol{0}_{T-p,h} \\ \boldsymbol{E}_1 & \boldsymbol{E}_2 \end{bmatrix},$$

where $\boldsymbol{E}_1 \in \mathbb{R}^{h \times T}$ and $\boldsymbol{E}_2 \in \mathbb{R}^{h \times h}$. For example, when $T = 4$, $h = 2$ and $p = 2$, thus, $\boldsymbol{E}_1 \in \mathbb{R}^{2 \times 4}$

[1] A MATLAB/GUN OBtave function for calculating $\widehat{\boldsymbol{z}}$ in 3.8 is shown in the Appendix.

and $\boldsymbol{E}_2 \in \mathbb{R}^{2\times 2}$:

$$\boldsymbol{D}_{4+2} = \left[\begin{array}{c:c} \boldsymbol{D}_4 & \mathbf{0}_{2,2} \\ \hdashline \boldsymbol{E}_1 & \boldsymbol{E}_2 \end{array}\right] = \left[\begin{array}{cccc:cc} 1 & -2 & 1 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 & 0 \\ \hdashline 0 & 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 0 & 1 & -2 & 1 \end{array}\right]. \tag{3.9}$$

Define $\widetilde{\boldsymbol{w}} = [\widetilde{z}_{T+1}, \ldots, \widetilde{z}_{T+h}]^\top$ by the requirement that:

$$\Delta^p \widetilde{z}_{T+j} = 0, \quad (j = 1, \ldots, h). \tag{3.10}$$

Then, concerning $\widehat{z}_1, \ldots, \widehat{z}_T, \widehat{z}_{T+1}, \ldots, \widehat{z}_{T+h}$, we have the results summarized in the following theorem:

Theorem 3.1.

$$\begin{cases} \widehat{z}_t = \widetilde{z}_t, & (t = 1, \ldots, T), \\ \widehat{z}_{T+j} = \widetilde{z}_{T+j} \text{ such that } \Delta^p \widetilde{z}_{T+j} = 0, & (j = 1, \ldots, h). \end{cases}$$

Proof of Theorem 3.1. By definition of $\widetilde{\boldsymbol{w}}$, we have $\boldsymbol{E}_1 \widetilde{\boldsymbol{z}} + \boldsymbol{E}_2 \widetilde{\boldsymbol{w}} = \mathbf{0}_{h,1}$, which leads to

$$\begin{aligned} \boldsymbol{D}_{T+h}^\top \boldsymbol{D}_{T+h} \begin{bmatrix} \widetilde{\boldsymbol{z}} \\ \widetilde{\boldsymbol{w}} \end{bmatrix} &= \begin{bmatrix} \boldsymbol{D}_T^\top & \boldsymbol{E}_1^\top \\ \mathbf{0}_{h,T-p} & \boldsymbol{E}_2^\top \end{bmatrix} \begin{bmatrix} \boldsymbol{D}_T & \mathbf{0}_{T-p,h} \\ \boldsymbol{E}_1 & \boldsymbol{E}_2 \end{bmatrix} \begin{bmatrix} \widetilde{\boldsymbol{z}} \\ \widetilde{\boldsymbol{w}} \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{D}_T^\top & \boldsymbol{E}_1^\top \\ \mathbf{0}_{h,T-p} & \boldsymbol{E}_2^\top \end{bmatrix} \begin{bmatrix} \boldsymbol{D}_T \widetilde{\boldsymbol{z}} \\ \mathbf{0}_{h,1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{D}_T^\top \boldsymbol{D}_T \widetilde{\boldsymbol{z}} \\ \mathbf{0}_{h,1} \end{bmatrix}. \end{aligned} \tag{3.11}$$

It follows from (3.7) and (3.11) that

$$(\boldsymbol{S}_T^\top \boldsymbol{S}_T + \lambda \boldsymbol{D}_{T+h}^\top \boldsymbol{D}_{T+h}) \begin{bmatrix} \widetilde{\boldsymbol{z}} \\ \widetilde{\boldsymbol{w}} \end{bmatrix} = \begin{bmatrix} \widetilde{\boldsymbol{z}} \\ \mathbf{0} \end{bmatrix} + \begin{bmatrix} \lambda \boldsymbol{D}_T^\top \boldsymbol{D}_T \widetilde{\boldsymbol{z}} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \boldsymbol{y} \\ \mathbf{0} \end{bmatrix} = \boldsymbol{S}_T^\top \boldsymbol{y}. \tag{3.12}$$

Premultiplying (3.12) by $(\boldsymbol{S}_T^\top \boldsymbol{S}_T + \lambda \boldsymbol{D}_{T+h}^\top \boldsymbol{D}_{T+h})^{-1}$, we finally obtain

$$\widehat{\boldsymbol{z}} = (\boldsymbol{S}_T^\top \boldsymbol{S}_T + \lambda \boldsymbol{D}_{T+h}^\top \boldsymbol{D}_{T+h})^{-1} \boldsymbol{S}_T^\top \boldsymbol{y} = \begin{bmatrix} \widetilde{\boldsymbol{z}} \\ \widetilde{\boldsymbol{w}} \end{bmatrix}. \tag{3.13}$$

□

Example 3.1. *As an illustration of above theorem, we give a simple numerical example. The case where $T = 5$, $p = 2$ and $h = 2$. Suppose that we obtained*

$$\widetilde{z}_1 = 2, \quad \Delta\widetilde{z}_2 = -1, \quad [\Delta^2\widetilde{z}_3, \Delta^2\widetilde{z}_4, \Delta^2\widetilde{z}_5]^\top = [3, 0, -1]^\top.$$

$\widetilde{z}_t$ for $t = 1, \dots, 5$ are explicitly $[\widetilde{z}_1, \widetilde{z}_2, \widetilde{z}_3, \widetilde{z}_4, \widetilde{z}_5]^\top = [2, 1, 3, 5, 6]^\top$. Then from the above theorem, in the case, $\widehat{z}_t$ for $t = 1, \dots, 5, 5+1, 5+2$ are as follows:

$$\begin{aligned} \widehat{\boldsymbol{z}} &= [\widetilde{z}_1, \widetilde{z}_2, \widetilde{z}_3, \widetilde{z}_4, \widetilde{z}_5, \widetilde{z}_{5+1}, \widetilde{z}_{5+2}]^\top \\ &= [2, 1, 3, 5, 6, 7, 8]^\top. \end{aligned}$$

The next theorem is a generalization of a result of Yamada (2017):

Theorem 3.2.

$$\lim_{\lambda \to \infty} \widehat{z}_t = \widehat{\beta}_0 + \widehat{\beta}_1 t + \cdots + \widehat{\beta}_{p-1} t^{p-1}, \quad (t = 1, \dots, T, T+1, \dots, T+h), \tag{3.14}$$

where $(\widehat{\beta}_0, \widehat{\beta}_1, \dots, \widehat{\beta}_{p-1}) = \arg\min_{\beta_0, \beta_1, \dots, \beta_{p-1}} \sum_{t=1}^T (y_t - \beta_0 - \beta_1 t - \cdots - \beta_{p-1} t^{p-1})^2$.

Proof of Theorem 3.2. Premultiplying (3.13) by $\boldsymbol{D}_{T+h}$ leads to

$$\boldsymbol{D}_{T+h} \widehat{\boldsymbol{z}} = \begin{bmatrix} \boldsymbol{D}_T & \mathbf{0}_{T-p,h} \\ \boldsymbol{E}_1 & \boldsymbol{E}_2 \end{bmatrix} \begin{bmatrix} \widetilde{\boldsymbol{z}} \\ \widetilde{\boldsymbol{w}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{D}_T \widetilde{\boldsymbol{z}} \\ \mathbf{0}_{h,1} \end{bmatrix}.$$

Since $\lim_{\lambda\to\infty}\widetilde{\boldsymbol{z}} = \boldsymbol{\Pi}_T(\boldsymbol{\Pi}_T^\top\boldsymbol{\Pi}_T)^{-1}\boldsymbol{\Pi}_T^\top\boldsymbol{y}$ and $\boldsymbol{D}_T\boldsymbol{\Pi}_T = \boldsymbol{0}$, we obtain

$$\boldsymbol{D}_{T+h}\lim_{\lambda\to\infty}\widehat{\boldsymbol{z}} = \begin{bmatrix}\boldsymbol{D}_T\lim_{\lambda\to\infty}\widetilde{\boldsymbol{z}}\\ \boldsymbol{0}_{h,1}\end{bmatrix} = \boldsymbol{0}_{T+h,1}, \tag{3.15}$$

which indicates $\lim_{\lambda\to\infty}\widehat{\boldsymbol{z}}$ is in the null space of $\boldsymbol{D}_{T+h}$. Since the null space of $\boldsymbol{D}_{T+h}$ and the column space of $\boldsymbol{\Pi}_{T+h}$ are equivalent, (3.15) implies that we obtain $\boldsymbol{\gamma}\in\mathbb{R}^p$ such that

$$\lim_{\lambda\to\infty}\widehat{\boldsymbol{z}} = \boldsymbol{\Pi}_{T+h}\boldsymbol{\gamma}. \tag{3.16}$$

Since $\boldsymbol{S}_T\lim_{\lambda\to\infty}\widehat{\boldsymbol{z}} = \boldsymbol{S}_T\boldsymbol{\Pi}_{T+h}\gamma = \boldsymbol{\Pi}_T\boldsymbol{\gamma}$ and $\boldsymbol{S}_T\lim_{\lambda\to\infty}\widehat{\boldsymbol{z}} = \lim_{\lambda\to\infty}\widetilde{\boldsymbol{z}} = \boldsymbol{\Pi}_T(\boldsymbol{\Pi}_T^\top\boldsymbol{\Pi}_T)^{-1}\boldsymbol{\Pi}_T^\top\boldsymbol{y}$, it follows that $\boldsymbol{\Pi}_T\{\boldsymbol{\gamma}-(\boldsymbol{\Pi}_T^\top\boldsymbol{\Pi}_T)^{-1}\boldsymbol{\Pi}_T^\top\boldsymbol{y}\} = \boldsymbol{0}_{h,1}$. Since $\boldsymbol{\Pi}_T$ is a full column rank matrix, we then obtain $\boldsymbol{\gamma} = (\boldsymbol{\Pi}_T^\top\boldsymbol{\Pi}_T)^{-1}\boldsymbol{\Pi}_T^\top\boldsymbol{y}$. Substituting this relation into (3.16), we finally obtain

$$\lim_{\lambda\to\infty}\widehat{\boldsymbol{z}} = \boldsymbol{\Pi}_{T+h}(\boldsymbol{\Pi}_T^\top\boldsymbol{\Pi}_T)^{-1}\boldsymbol{\Pi}_T^\top\boldsymbol{y}. \tag{3.17}$$

□

Finally, we give some related results. Since $\boldsymbol{D}_{T+h}\boldsymbol{\Pi}_{T+h} = \boldsymbol{0}_{T-p+h,p}$, it follows that

$$(\boldsymbol{S}_T^\top\boldsymbol{S}_T + \lambda\boldsymbol{D}_{T+h}^\top\boldsymbol{D}_{T+h})\boldsymbol{\Pi}_{T+h} = \boldsymbol{S}_T^\top\boldsymbol{S}_T\boldsymbol{\Pi}_{T+h} = \boldsymbol{S}_T^\top\boldsymbol{\Pi}_T.$$

Premultiplying the above equation by $(\boldsymbol{S}_T^\top\boldsymbol{S}_T + \lambda\boldsymbol{D}_{T+h}^\top\boldsymbol{D}_{T+h})^{-1}$ yields

$$\boldsymbol{\Pi}_{T+h} = (\boldsymbol{S}_T^\top\boldsymbol{S}_T + \lambda\boldsymbol{D}_{T+h}^\top\boldsymbol{D}_{T+h})^{-1}\boldsymbol{S}_T^\top\boldsymbol{\Pi}_T. \tag{3.18}$$

Example 3.2. *We give a numberical example as an illustration of* (3.18). *Consider the case where*

$T=4$, $h=1$ and $p=2$, it follows that

$$\begin{bmatrix}1 & 1\\ 1 & 2\\ 1 & 3\\ 1 & 4\\ 1 & 5\end{bmatrix} = (\boldsymbol{S}_4^\top \boldsymbol{S}_4 + \lambda \boldsymbol{D}_5^\top \boldsymbol{D}_5)^{-1} \boldsymbol{S}_4^\top \begin{bmatrix}1 & 1\\ 1 & 2\\ 1 & 3\\ 1 & 4\end{bmatrix}.$$

From (3.18), we observe that the row sums of $(\boldsymbol{S}_T^\top \boldsymbol{S}_T + \lambda \boldsymbol{D}_{T+h}^\top \boldsymbol{D}_{T+h})^{-1} \boldsymbol{S}_T^\top$ always equal unity. Let $\widetilde{\boldsymbol{\tau}} = \boldsymbol{\Pi}_T(\boldsymbol{\Pi}_T^\top \boldsymbol{\Pi}_T)^{-1}\boldsymbol{\Pi}_T^\top \boldsymbol{y}$. Premultiplying $\boldsymbol{y} = \widetilde{\boldsymbol{\tau}} + (\boldsymbol{y} - \widetilde{\boldsymbol{\tau}})$ by $(\boldsymbol{S}_T^\top \boldsymbol{S}_T + \lambda \boldsymbol{D}_{T+h}^\top \boldsymbol{D}_{T+h})^{-1} \boldsymbol{S}_T^\top$, we obtain

$$\begin{aligned}\widehat{\boldsymbol{z}} &= (\boldsymbol{S}_T^\top \boldsymbol{S}_T + \lambda \boldsymbol{D}_{T+h}^\top \boldsymbol{D}_{T+h})^{-1} \boldsymbol{S}_T^\top \widetilde{\boldsymbol{\tau}} + (\boldsymbol{S}_T^\top \boldsymbol{S}_T + \lambda \boldsymbol{D}_{T+h}^\top \boldsymbol{D}_{T+h})^{-1} \boldsymbol{S}_T^\top (\boldsymbol{y} - \widetilde{\boldsymbol{\tau}})\\ &= \boldsymbol{\Pi}_{T+h}(\boldsymbol{\Pi}_T^\top \boldsymbol{\Pi}_T)^{-1}\boldsymbol{\Pi}_T^\top \boldsymbol{y} + (\boldsymbol{S}_T^\top \boldsymbol{S}_T + \lambda \boldsymbol{D}_{T+h}^\top \boldsymbol{D}_{T+h})^{-1} \boldsymbol{S}_T^\top (\boldsymbol{y} - \widetilde{\boldsymbol{\tau}}). \qquad (3.19)\end{aligned}$$

(3.18) and (3.19) are generalizations of Eq. (2) and Eq. (3) respectively of Yamada (2018). We note that $\boldsymbol{\Pi}_{T+h}(\boldsymbol{\Pi}_T^\top \boldsymbol{\Pi}_T)^{-1}\boldsymbol{\Pi}_T^\top \boldsymbol{y}$ in (3.19) appears in (3.17), and from this we find that

$$\lim_{\lambda\to\infty} (\boldsymbol{S}_T^\top \boldsymbol{S}_T + \lambda \boldsymbol{D}_{T+h}^\top \boldsymbol{D}_{T+h})^{-1} \boldsymbol{S}_T^\top (\boldsymbol{y} - \widetilde{\boldsymbol{\tau}}) = \boldsymbol{0}_{T+h,1}. \qquad (3.20)$$

3.3 Two other penalized least squares problems

In this section, we introduce two other penalized least squares problems and show that they give the same solutions as (3.6). Let $y_{T+j} = \widetilde{z}_{T+j}$ for $j = 1, \ldots, h$. Consider the following two penalized least squares problems:

$$(\widehat{z}_1^{(a)}, \ldots, \widehat{z}_T^{(a)}, \widehat{z}_{T+1}^{(a)}, \ldots, \widehat{z}_{T+h}^{(a)}) = \underset{z_1,\ldots,z_T,z_{T+1},\ldots,z_{T+h}\in\mathbb{R}}{\arg\min} \sum_{t=1}^{T+h}(y_t - z_t)^2 + \lambda \sum_{t=p+1}^{T+h}(\Delta^p z_t)^2, \qquad (3.21)$$

$$(\widehat{z}_1^{(b)}, \ldots, \widehat{z}_T^{(b)}, \widehat{z}_{T+1}^{(b)}, \ldots, \widehat{z}_{T+h}^{(b)}) = \underset{z_1,\ldots,z_T,z_{T+1},\ldots,z_{T+h}\in\mathbb{R}}{\arg\min} \sum_{t=1}^{T+h}(y_t - z_t)^2 + \lambda \sum_{t=p+1}^{T}(\Delta^p z_t)^2, \qquad (3.22)$$

Letting $\widehat{\boldsymbol{z}}^{(i)} = [\widehat{z}_1^{(i)}, \ldots, \widehat{z}_T^{(i)}, \widehat{z}_{T+1}^{(i)}, \ldots, \widehat{z}_{T+h}^{(i)}]^\top$ for $i = a, b$, we show that $\widehat{\boldsymbol{z}}^{(i)}$ for $i = a, b$ are then expressed explicitly as

$$\widehat{\boldsymbol{z}}^{(a)} = (\boldsymbol{I}_{T+h} + \lambda \boldsymbol{D}_{T+h}^\top \boldsymbol{D}_{T+h})^{-1} \begin{bmatrix} \boldsymbol{y} \\ \widetilde{\boldsymbol{w}} \end{bmatrix}, \tag{3.23}$$

$$\widehat{\boldsymbol{z}}^{(b)} = (\boldsymbol{I}_{T+h} + \lambda \boldsymbol{D}_{T+h}^\top \boldsymbol{S}_{T-p}^\top \boldsymbol{S}_{T-p} \boldsymbol{D}_{T+h})^{-1} \begin{bmatrix} \boldsymbol{y} \\ \widetilde{\boldsymbol{w}} \end{bmatrix}, \tag{3.24}$$

where we recall that $\widetilde{\boldsymbol{w}} = [\widetilde{z}_{T+1}, \ldots, \widetilde{z}_{T+h}]^\top$ is defined by (3.10) and is obtainable as a part of the solution of the modified WH graduation, (3.6).

Theorem 3.3. *For* $i = a, b$,

$$\begin{cases} \widehat{z}_t^{(i)} = \widetilde{z}_t, & (t = 1, \ldots, T), \\ \widehat{z}_{T+j}^{(i)} = \widetilde{z}_{T+j} \text{ such that } \Delta^p \widetilde{z}_{T+j} = 0, & (j = 1, \ldots, h). \end{cases}$$

Proof of Theorem 3.3 ($i = a$). From (3.11), we obtain

$$(\boldsymbol{I}_{T+h} + \lambda \boldsymbol{D}_{T+h}^\top \boldsymbol{D}_{T+h}) \begin{bmatrix} \widetilde{\boldsymbol{z}} \\ \widetilde{\boldsymbol{w}} \end{bmatrix} = \begin{bmatrix} \widetilde{\boldsymbol{z}} \\ \widetilde{\boldsymbol{w}} \end{bmatrix} + \begin{bmatrix} \lambda \boldsymbol{D}_T^\top \boldsymbol{D}_T \widetilde{\boldsymbol{z}} \\ \boldsymbol{0}_{h,1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{y} \\ \widetilde{\boldsymbol{w}} \end{bmatrix}. \tag{3.25}$$

Premultiplying (3.25) by $(\boldsymbol{I}_{T+h} + \lambda \boldsymbol{D}_{T+h}^\top \boldsymbol{D}_{T+h})^{-1}$, it follows that

$$\widehat{\boldsymbol{z}}^{(a)} = (\boldsymbol{I}_{T+h} + \lambda \boldsymbol{D}_{T+h}^\top \boldsymbol{D}_{T+h})^{-1} \begin{bmatrix} \boldsymbol{y} \\ \widetilde{\boldsymbol{w}} \end{bmatrix} = \begin{bmatrix} \widetilde{\boldsymbol{z}} \\ \widetilde{\boldsymbol{w}} \end{bmatrix}.$$

□

Proof of Theorem 3.3 ($i = b$). Since

$$\boldsymbol{D}_{T+h}^\top \boldsymbol{S}_{T-p}^\top \boldsymbol{S}_{T-p} \boldsymbol{D}_{T+h} \begin{bmatrix} \widetilde{\boldsymbol{z}} \\ \widetilde{\boldsymbol{w}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{D}_T^\top & \boldsymbol{E}_1^\top \\ \boldsymbol{0}_{h,T-p} & \boldsymbol{E}_2^\top \end{bmatrix} \begin{bmatrix} \boldsymbol{I}_{T-p} \\ \boldsymbol{0}_{h,T-p} \end{bmatrix} [\boldsymbol{I}_{T-p}, \boldsymbol{0}_{T-p,h}] \begin{bmatrix} \boldsymbol{D}_T & \boldsymbol{0}_{T-p,h} \\ \boldsymbol{E}_1 & \boldsymbol{E}_2 \end{bmatrix} \begin{bmatrix} \widetilde{\boldsymbol{z}} \\ \widetilde{\boldsymbol{w}} \end{bmatrix}$$

$$= \begin{bmatrix} \boldsymbol{D}_T^\top \\ \boldsymbol{0}_{h,T-p} \end{bmatrix} [\boldsymbol{D}_T, \boldsymbol{0}_{T-p,h}] \begin{bmatrix} \widetilde{\boldsymbol{z}} \\ \widetilde{\boldsymbol{w}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{D}_T^\top \boldsymbol{D}_T \widetilde{\boldsymbol{z}} \\ \boldsymbol{0}_{h,1} \end{bmatrix},$$

it follows that

$$(\boldsymbol{I}_{T+h} + \lambda \boldsymbol{D}_{T+h}^\top \boldsymbol{S}_{T-p}^\top \boldsymbol{S}_{T-p} \boldsymbol{D}_{T+h}) \begin{bmatrix} \widetilde{\boldsymbol{z}} \\ \widetilde{\boldsymbol{w}} \end{bmatrix} = \begin{bmatrix} \widetilde{\boldsymbol{z}} \\ \widetilde{\boldsymbol{w}} \end{bmatrix} + \begin{bmatrix} \lambda \boldsymbol{D}_T^\top \boldsymbol{D}_T \widetilde{\boldsymbol{z}} \\ \boldsymbol{0}_{h,1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{y} \\ \widetilde{\boldsymbol{w}} \end{bmatrix}. \tag{3.26}$$

Premultiplying (3.26) by $(\boldsymbol{I}_{T+h} + \lambda \boldsymbol{D}_{T+h}^\top \boldsymbol{S}_{T-p}^\top \boldsymbol{S}_{T-p} \boldsymbol{D}_{T+h})^{-1}$ yields

$$\widehat{\boldsymbol{z}}^{(b)} = (\boldsymbol{I}_{T+h} + \lambda \boldsymbol{D}_{T+h}^\top \boldsymbol{S}_{T-p}^\top \boldsymbol{S}_{T-p} \boldsymbol{D}_{T+h})^{-1} \begin{bmatrix} \boldsymbol{y} \\ \widetilde{\boldsymbol{w}} \end{bmatrix} = \begin{bmatrix} \widetilde{\boldsymbol{z}} \\ \widetilde{\boldsymbol{w}} \end{bmatrix}.$$

□

Example 3.3. *As an illustration of above theorem, we give a numerical example. The case where $T = 6$, $p = 3$ and $h = 3$. Suppose that we obtained*

$$\widetilde{z}_1 = 1, \quad \Delta \widetilde{z}_2 = -2, \quad \Delta^2 \widetilde{z}_3 = 1, \quad [\Delta^3 \widetilde{z}_4, \Delta^3 \widetilde{z}_5, \Delta^3 \widetilde{z}_6]^\top = [4, -11, 7]^\top.$$

by applying polynomial trend filtering of order 3 to a T-dimensional time series data. $\widetilde{z}_1$ for $t = 1, \ldots, 6$ are explicitly $[\widetilde{z}_1, \widetilde{z}_2, \widetilde{z}_3, \widetilde{z}_4, \widetilde{z}_5, \widetilde{z}_6]^\top = [1, -1, -2, 2, 0, -1]^\top$. Then from the above theorem, in the case, $\widehat{z}_t^{(i)}$ for $i = a, b$ and $t = 1, \ldots, 6, 6+1, 6+2, 6+3$ are as follows:

$$\begin{aligned} \widehat{z}_t^{(i)} &= [\widehat{z}_1^{(i)}, \widehat{z}_2^{(i)}, \widehat{z}_3^{(i)}, \widehat{z}_4^{(i)}, \widehat{z}_5^{(i)}, \widehat{z}_6^{(i)}, \widehat{z}_{6+1}^{(i)}, \widehat{z}_{6+2}^{(i)}, \widehat{z}_{6+3}^{(i)}]^\top \\ &= [1, -1, -2, 2, 0, -1, -1, 0, 2]^\top. \end{aligned}$$

Remark 3.4. An argument similar to that in Theorem 3.3 ($i = a$) is given by Mohr (2005, p. 20).

From Theorems 3.1, 3.2, and 3.3, we immediately obtain the following theorem.

Theorem 3.5. *For $i = a, b$,*

$$\lim_{\lambda \to \infty} \widehat{z}_t^{(i)} = \widehat{\beta}_0 + \widehat{\beta}_1 t + \cdots + \widehat{\beta}_{p-1} t^{p-1}, \quad (t = 1, \ldots, T, T+1, \ldots, T+h), \tag{3.27}$$

where $(\widehat{\beta}_0, \widehat{\beta}_1, \dots, \widehat{\beta}_{p-1}) = \arg\min_{\beta_0,\beta_1,\dots,\beta_{p-1}} \sum_{t=1}^{T}(y_t - \beta_0 - \beta_1 t - \cdots - \beta_{p-1}t^{p-1})^2$.

Example 3.4. *We give a numberical example for the case* $\boldsymbol{y} = [1, 2, -1, 3, -2]$*,* $T = 5$*,* $h = 2$ *and* $p = 3$*. According to* $(\widehat{\beta}_0, \widehat{\beta}_1, \dots, \widehat{\beta}_{p-1}) = \arg\min_{\beta_0,\beta_1,\dots,\beta_{p-1}} \sum_{t=1}^{T}(y_t - \beta_0 - \beta_1 t - \cdots - \beta_{p-1}t^{p-1})^2$*. suppose that we have a linear regression* $\boldsymbol{y} = \boldsymbol{\Pi}_5\boldsymbol{\beta}_i$ *for* $i = 1, 2, 3$*, then we rewrite it in matrix form as follows*

$$\boldsymbol{y} = \boldsymbol{\Pi}_5\boldsymbol{\beta}_i = \begin{bmatrix} 1 \\ 2 \\ -1 \\ 3 \\ -2 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \\ 1 & 5 & 25 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}.$$

we obtain $\widehat{\boldsymbol{\beta}} = [\beta_0, \beta_1, \beta_2]^\top = [-0.4000, 1.64286, -0.35714]^\top$.

From the above theorem,

$$\lim_{\lambda\to\infty} \widehat{z}_t^{(i)} = \widehat{\beta}_0 + \widehat{\beta}_1 t + \widehat{\beta}_2 t^2.$$

where $t = 1, \dots, 5, 5+1, 5+2$ *and* $i = a, b$*.*

we obtain that

$$\lim_{\lambda\to\infty} \widehat{z}_t^{(i)} = \boldsymbol{\Pi}_7\boldsymbol{\beta}_i = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 2^2 \\ \vdots & \vdots & \vdots \\ 1 & 5 & 5^2 \\ 1 & 6 & 6^2 \\ 1 & 7 & 7^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

$$= [0.88571, 1.45714, 1.31429, 0.45714, -1.11429, -3.40000, -6.40000]^\top.$$

3.4 Concluding remarks

This chapter presents a modified Whittaker–Henderson (WH) Method of Graduation. After giving a closed-form solution, we show that it is of practical use because it provides not only a smoothed series identical to that of the WH graduation, but also an extrapolation beyond the sample limit of current data. In addition, we introduce two other penalized least squares problems and show that they provide the same results as those of the modified WH graduation.

3.5 Appendix

3.5.1 MATLAB/GNU Octave function for calculating $\widehat{z}$ in (3.8)

```
function zhat=mWHgraduation(y,lamda,p,h)
% y: T-dimensional column vector
% lamda: positive constant
% zhat : (T+h)-dimensional column vector
   T=length(y);
   S=[eye(T),zeros(T, h)];
   D=diff(eye(T+h),p);
   zhat=inv(S'*S+lambda*D'*D)*(S'*y);
end
```

3.6 References

1. Bohlmann, G., 1899, Ein Ausgleichungsproblem, *Nachrichten von der Gesellschaft der Wissenschaften zu Gottingen, Mathematisch-Physikalische Klasse*, 260–271.

2. Cornea-Madeira, A., 2017, The explicit formula for the Hodrick–Prescott filter in finite sample, *Review of Economics and Statistics*, 99, 2, 314–318.

3. De Jong, R. M. and N. Sakarya, 2016, The econometrics of the Hodrick–Prescott filter, *Review of Economics and Statistics*, 98, 2, 310–317.

4. Hodrick, R. J. and E. C. Prescott, 1997, Postwar U.S. business cycles: An empirical investigation, *Journal of Money, Credit and Banking*, 29, 1, 1–16.

5. Nocon, A. S. and W. F. Scott, 2012, An extension of the Whittaker–Henderson method of graduation, *Scandinavian Actuarial Journal*, 2012, 1, 70–79.

6. Pesaran, M. H., 1973, Exact maximum likelihood estimation of a regression equation with a first-order moving-average error, *Review of Economic Studies*, 40, 4, 529–535.

7. Phillips P. C. B., 2010, Two New Zealand pioneer econometricians, *New Zealand Economic Papers*, 44, 1, 1–26.

8. Strang, G. and S. MacNamara, 2014, Functions of difference matrices are Toeplitz plus Hankel, *SIAM Review*, 56, 3, 525–546.

9. Weinert, H. L., 2007, Efficient computation for Whittaker–Henderson smoothing, *Computational Statistics and Data Analysis*, 52, 2, 959–974.

10. Yamada, H., 2015, Ridge regression representations of the generalized Hodrick–Prescott filter, *Journal of the Japan Statistical Society*, 45, 2, 121–128.

11. Yamada, H., 2018, Several least squares problems related to the Hodrick–Prescott filtering, *Communications in Statistics – Theory and Methods*, 47, 5, 1022–1027.

12. Yamada, H. and R. Du, 2019, A modification of the Whittaker–Henderson method of graduation, *Communications in Statistics – Theory and Methods*, 48, 15, 3795–3800.

Chapter 4

Some Results on ℓ_1 Polynomial Trend Filtering

This chapter is based on a previously published article: Yamada and Du (2018).

4.1 Introduction

The ℓ_1-norm penalized least-squares problem, defined as:

$$\min_{x_1,\ldots,x_T} \sum_{t=1}^{T}(y_t - x_t)^2 + \lambda \sum_{t=3}^{T}|\Delta^2 x_t|, \tag{4.1}$$

where $y_1, \ldots, y_T$ are observed time-series data, was developed by Kim et al. (2009), who called it ℓ_1 trend filtering.[1] Here, $\lambda > 0$ is a tuning parameter and Δ denotes the backward difference operator such that $\Delta x_t = x_t - x_{t-1}$. Accordingly, $\Delta^2 x_t = \Delta(\Delta x_t) = x_t - 2x_{t-1} + x_{t-2}$. Recall that $\sum_{t=3}^{T}|\Delta^2 x_t|$ in (4.1) is ℓ_1-norm of $[\Delta^2 x_3, \ldots, \Delta^2 x_T]^\top$. Unlike Hodrick–Prescott (HP) (1997) filtering, which is defined as the following squared ℓ_2-norm penalized least-squares problem:

$$\min_{x_1,\ldots,x_T} \sum_{t=1}^{T}(y_t - x_t)^2 + \psi \sum_{t=3}^{T}(\Delta^2 x_t)^2, \tag{4.2}$$

[1] ℓ_1 trend filtering is supported in several standard software packages such as MATLAB, R, Python, and EViews.

where $\psi > 0$ is a smoothing/tuning parameter, the solution of ℓ_1 trend filtering becomes a continuous piecewise linear trend. The relationship between HP filtering and ℓ_1 trend filtering corresponds to that between ridge regression of Hoerl and Kennard (1970) and Lasso (least absolute shrinkage and selection operator) regression of Tibshirani (1996)/BPDN (basis pursuit denoising) of Chen et al. (1998). Econometric applications of ℓ_1 trend filtering include Yamada and Jin (2013), Yamada and Yoon (2014), Winkelried (2016), and Yamada (2017a).

It has been well-known that HP filtering is a form of the Whittaker–Henderson (WH) method of graduation, which is defined as:

$$\min_{x_1,\ldots,x_T} \sum_{t=1}^{T}(y_t - x_t)^2 + \psi \sum_{t=p+1}^{T} (\Delta^p x_t)^2. \tag{4.3}$$

For historical surveys of WH filtering, see Weinert (2007), Phillips (2010), and Nocon and Scott (2012). Likewise, as shown in Kim et al. (2009), Tibshirani and Taylor (2011), and Tibshirani (2014), ℓ_1 trend filtering may be generalized as:

$$\min_{x_1,\ldots,x_T} \sum_{t=1}^{T}(y_t - x_t)^2 + \lambda \sum_{t=p+1}^{T} |\Delta^p x_t|. \tag{4.4}$$

We refer to it as ℓ_1 polynomial trend filtering.[2] This filtering method is promising because it enables

[2](4.4) where $p = 1$ has been known as total variation denoising in signal processing, which may be regarded as a form of the fused Lasso by Tibshirani et al. (2005). Harchaoui and Lévy-Leduc (2010) proposed using the filtering to detect multiple change points. (4.4) may be regarded as a form of the generalized Lasso by Tibshirani and Taylor (2011). In addition, we note that there exist some pioneering works on the filtering that uses the ℓ_1-norm penalty. Miller (1946, Sec. 1.7) mentioned that $\sum_{t=p+1}^{T}|\Delta^p x_t|$ could be an alternative measure of smoothness to $\sum_{t=p+1}^{T}(\Delta^p x_t)^2$, Schuette (1978) introduced a filtering, defined as:

$$\min_{x_1,\ldots,x_T} \sum_{t=1}^{T}|y_t - x_t| + \lambda \sum_{t=p+1}^{T} |\Delta^p x_t|,$$

and Koenker et al. (1994) presented ℓ_1-norm penalized quantile smoothing spline. Incidentally, Schuette (1978) and Koenker et al. (1994) motivate us to consider a penalized quantile regression that is obtainable by replacing the quadratic loss function in (4.4) by the check loss function:

$$\min_{x_1,\ldots,x_T} \sum_{t=1}^{T}\rho_\tau(y_t - x_t) + \lambda \sum_{t=p+1}^{T} |\Delta^p x_t|,$$

where, letting $\tau \in (0, 1)$,

$$\rho_\tau(u) = \begin{cases} \tau|u| & (u \geq 0), \\ (1-\tau)|u| & (u < 0), \end{cases}$$

which is suggested by Kim et al. (2009, Sec. 7.3).

us to estimate a piecewise $(p-1)$-th order polynomial trend of a univariate economic time series without prespecifying the number and location of knots. For more details, see Yamada (2017b).

Let $\widehat{x}_1, \ldots, \widehat{x}_T$ denote the solution of (4.3) and define $\widehat{x}_{T+1}, \ldots, \widehat{x}_{T+h}$, where h denotes the length of extrapolation by:

$$\Delta^p \widehat{x}_{T+j} = 0, \quad (j = 1, \ldots, h). \tag{4.5}$$

Recently, Yamada and Du (2018) introduced the following three modifications of the WH method of graduation:[3]

$$\text{(a)} \min_{x_1,\ldots,x_{T+h}} \sum_{t=1}^{T} (y_t - x_t)^2 + \psi \sum_{t=p+1}^{T+h} (\Delta^p x_t)^2, \tag{4.6}$$

$$\text{(b)} \min_{x_1,\ldots,x_{T+h}} \sum_{t=1}^{T+h} (y_t - x_t)^2 + \psi \sum_{t=p+1}^{T+h} (\Delta^p x_t)^2, \tag{4.7}$$

$$\text{(c)} \min_{x_1,\ldots,x_{T+h}} \sum_{t=1}^{T+h} (y_t - x_t)^2 + \psi \sum_{t=p+1}^{T} (\Delta^p x_t)^2, \tag{4.8}$$

where $y_{T+j} = \widehat{x}_{T+j}$ for $j = 1, \ldots, h$. Denote the solution of (a), (b), and (c) by $\widehat{x}_t^{(i)}$ for $i = \text{a}, \text{b}, \text{c}$ and $t = 1, \ldots, T+h$. Yamada and Du (2018) showed that, for $i = \text{a}, \text{b}, \text{c}$ and $t = 1, \ldots, T+h$, it follows that:

$$\widehat{x}_t^{(i)} = \widehat{x}_t. \tag{4.9}$$

Among the above results, $\widehat{x}_t^{(\text{a})} = \widehat{x}_t$ is of practical use because it provides not only a smoothed series identical to that of the WH graduation, but also an extrapolation beyond the sample limit of current data. Also, $\widehat{x}_t^{(\text{b})} = \widehat{x}_t$ is of interest because it shows that $\widehat{x}_{T+1}, \ldots, \widehat{x}_{T+h}$ based on (4.5) are useless to reduce the end-point problem of the WH graduation.[4] In addition, Yamada and Du (2018) proved that, for $i = \text{a}, \text{b}, \text{c}$ and $t = 1, \ldots, T+h$:

$$\lim_{\psi \to \infty} \widehat{x}_t^{(i)} = \widehat{\beta}_0 t^0 + \cdots + \widehat{\beta}_{p-1} t^{p-1}, \tag{4.10}$$

[3] See also Yamada (2017c).

[4] An argument similar to this is given by Mohr (2005, p. 20).

where $(\widehat{\beta}_0, \ldots, \widehat{\beta}_{p-1}) = \arg\min_{\beta_0,\ldots,\beta_{p-1}} \sum_{t=1}^{T}(y_t - \beta_0 t^0 - \cdots - \beta_{p-1}t^{p-1})^2$.

In this paper, we present three modifications of ℓ_1 polynomial trend filtering and show that they provide not only identical trend estimates as ℓ_1 polynomial trend filtering, but also extrapolations of the trend beyond both sample limits. In addition, we show some other results on the modified filtering. We also provide a MATLAB function for calculating the solution of one of the modified filtering methods.

The chapter is organized as follows. In Section 4.2, we present three modifications of ℓ_1 polynomial trend filtering. In Section 4.3, we state the main results of the paper. In Section 4.4, we give some remarks on the results provided in Section 3. Section 4.5 provides some concluding remarks.

Notations Let $\boldsymbol{y} = [y_1, \ldots, y_T]^\top$ and $\boldsymbol{I}_T$ be the $T \times T$ identity matrix. For an n-dimensional column vector, $\boldsymbol{\eta} = [\eta_1, \ldots, \eta_n]^\top$, $\|\boldsymbol{\eta}\|_1 = \sum_{i=1}^{n}|\eta_i|$, $\|\boldsymbol{\eta}\|_2^2 = \sum_{i=1}^{n} \eta_i^2$, and $\|\boldsymbol{\eta}\|_\infty = \max(|\eta_1|, \ldots, |\eta_n|)$. $\boldsymbol{D}_n$ is the $(n-p) \times n$ p-th order difference matrix such that $\boldsymbol{D}_n\boldsymbol{\eta} = [\Delta^p\eta_{p+1}, \ldots, \Delta^p\eta_n]^\top$. We denote $\boldsymbol{D}_T$ by $\boldsymbol{D}$. $\boldsymbol{\Pi}_{g+T+h}$ is a $(g+T+h) \times p$ Vandermonde matrix, defined by

$$\boldsymbol{\Pi}_{g+T+h} = \begin{bmatrix} (1-g)^0 & (1-g)^1 & \cdots & (1-g)^{p-1} \\ \vdots & \vdots & & \vdots \\ 1^0 & 1^1 & \cdots & 1^{p-1} \\ \vdots & \vdots & & \vdots \\ T^0 & T^1 & \cdots & T^{p-1} \\ \vdots & \vdots & & \vdots \\ (T+h)^0 & (T+h)^1 & \cdots & (T+h)^{p-1} \end{bmatrix},$$

and we denote $\boldsymbol{\Pi}_{0+T+0}$, which is a $T \times p$ matrix, by $\boldsymbol{\Pi}$.

4.2 Three modifications of ℓ_1 polynomial trend filtering

Let $\widetilde{x}_1, \ldots, \widetilde{x}_T$ denote the solution of (4.4) and define $\widetilde{x}_{1-g}, \ldots, \widetilde{x}_{1-1}$ and $\widetilde{x}_{T+1}, \ldots, \widetilde{x}_{T+h}$, where g and h denote the length of extrapolations:

$$\Delta^p \widetilde{x}_{p+1-i} = 0, \quad (i = 1, \ldots, g), \tag{4.11}$$

$$\Delta^p \widetilde{x}_{T+j} = 0, \quad (j = 1, \ldots, h). \tag{4.12}$$

For example, $\widetilde{x}_{T+1}, \ldots, \widetilde{x}_{T+h}$, defined by (4.12) for $p = 1, 2, 3$, are explicitly expressed as follows:

$$(p = 1) \quad \widetilde{x}_{T+j} = \widetilde{x}_T, \quad (j = 1, \ldots, h), \tag{4.13}$$

$$(p = 2) \quad \widetilde{x}_{T+j} = \widetilde{x}_T + j(\Delta\widetilde{x}_T), \quad (j = 1, \ldots, h), \tag{4.14}$$

$$(p = 3) \quad \widetilde{x}_{T+j} = \widetilde{x}_T + j(\Delta\widetilde{x}_T) + \frac{j(j+1)}{2}(\Delta^2\widetilde{x}_T), \quad (j = 1, \ldots, h). \tag{4.15}$$

For a proof of (4.15), see the Appendix.

Consider the following three modifications of ℓ_1 polynomial trend filtering:

$$\text{(d)} \min_{x_{1-g},\ldots,x_{T+h}} \sum_{t=1}^{T}(y_t - x_t)^2 + \lambda \sum_{t=p+1-g}^{T+h} |\Delta^p x_t|, \tag{4.16}$$

$$\text{(e)} \min_{x_{1-g},\ldots,x_{T+h}} \sum_{t=1-g}^{T+h}(y_t - x_t)^2 + \lambda \sum_{t=p+1-g}^{T+h} |\Delta^p x_t|, \tag{4.17}$$

$$\text{(f)} \min_{x_{1-g},\ldots,x_{T+h}} \sum_{t=1-g}^{T+h}(y_t - x_t)^2 + \lambda \sum_{t=p+1}^{T} |\Delta^p x_t|, \tag{4.18}$$

where $y_{1-i} = \widetilde{x}_{1-i}$ for $i = 1, \ldots, g$ and $y_{T+j} = \widetilde{x}_{T+j}$ for $j = 1, \ldots, h$. Note that (4.16) is equivalent to ℓ_1 polynomial trend filtering if $g = h = 0$. We denote the solution of (d), (e), and (f) by $\widetilde{x}_t^{(i)}$ for $i = \mathrm{d}, \mathrm{e}, \mathrm{f}$ and $t = 1-g, \ldots, T+h$.

Among (4.16), (4.17), and (4.18), the objective function of (4.16) may be represented in matrix notation as:[5]

$$\|\boldsymbol{y} - \boldsymbol{S}\boldsymbol{x}_{g+T+h}\|_2^2 + \lambda\|\boldsymbol{D}_{g+T+h}\boldsymbol{x}_{g+T+h}\|_1, \tag{4.19}$$

[5] For an n-dimensional column vector, $\boldsymbol{\eta} = [\eta_1, \ldots, \eta_n]^\top$, $\|\boldsymbol{\eta}\|_1 = \sum_{i=1}^n |\eta_i|$, $\|\boldsymbol{\eta}\|_2^2 = \sum_{i=1}^n \eta_i^2$, and $\|\boldsymbol{\eta}\|_\infty = \max(|\eta_1|, \ldots, |\eta_n|)$.

where $\boldsymbol{y} = [y_1, \ldots, y_T]^\top$ and $\boldsymbol{S} = [\boldsymbol{0}, \boldsymbol{I}_T, \boldsymbol{0}]$ is a $T \times (g+T+h)$ matrix and $\boldsymbol{x}_{g+T+h}$ is a $(g+T+h)$-dimensional column vector. $\boldsymbol{D}_{T+h+g}$ is the $(T+h+g-p) \times (T+h+g)$ p-th order difference matrix and we denote $\boldsymbol{D}_T$ by $\boldsymbol{D}$. Let $\widetilde{\boldsymbol{x}}_{g+T+h}^{(\mathrm{d})} = [\widetilde{\boldsymbol{x}}_g^{(\mathrm{d})\top}, \widetilde{\boldsymbol{x}}^{(\mathrm{d})\top}, \widetilde{\boldsymbol{x}}_h^{(\mathrm{d})\top}]^\top$, where $\widetilde{\boldsymbol{x}}_g^{(\mathrm{d})} = [\widetilde{x}_{1-g}^{(\mathrm{d})}, \ldots, \widetilde{x}_{1-1}^{(\mathrm{d})}]^\top$, $\widetilde{\boldsymbol{x}}^{(\mathrm{d})} = [\widetilde{x}_1^{(\mathrm{d})}, \ldots, \widetilde{x}_T^{(\mathrm{d})}]^\top$, and $\widetilde{\boldsymbol{x}}_h^{(\mathrm{d})} = [\widetilde{x}_{T+1}^{(\mathrm{d})}, \ldots, \widetilde{x}_{T+h}^{(\mathrm{d})}]^\top$. The MATLAB function for calculating $\widetilde{\boldsymbol{x}}_g^{(\mathrm{d})}$, $\widetilde{\boldsymbol{x}}^{(\mathrm{d})}$, and $\widetilde{\boldsymbol{x}}_h^{(\mathrm{d})}$, which depends on CVX developed by Grant and Boyd (2013), is as follows:

```
function [x_g,x,x_h]=m_l1_pt_filtering(y,lambda,p,g,h)
% y: T-dimensional column vector
% lambda: positive real number
% p, g, h: positive integer
% x_g: g-dimensional column vector
% x: T-dimensional column vector
% x_h: h-dimensional column vector
   T=length(y);
   S=[sparse(T,g),speye(T),sparse(T,h)];
   D=diff(speye(g+T+h),p);
   cvx_begin
      variables z(g+T+h)
      minimize(sum((y-S*z).^2)+lambda*norm(D*z,1))
   cvx_end
   x_g=z(1:g); x=z(g+1:g+T); x_h=z(g+T+1:g+T+h);
end
```

4.3 Main results

Theorem 4.1. *Denote the solution of (d), (e), and (f) by $\widetilde{x}_t^{(i)}$ for $i = \mathrm{d}, \mathrm{e}, \mathrm{f}$. For $i = \mathrm{d}, \mathrm{e}, \mathrm{f}$, and $t = 1-g, \ldots, T+h$, it follows that:*

$$\widetilde{x}_t^{(i)} = \widetilde{x}_t, \tag{4.20}$$

where $\widetilde{x}_1, \ldots, \widetilde{x}_T$ are the solution of (4.4) *and $\widetilde{x}_{1-g}, \ldots, \widetilde{x}_{1-1}$ and $\widetilde{x}_{T+1}, \ldots, \widetilde{x}_{T+h}$ are defined by* (4.11) *and* (4.12).

Proof. Because the objective function of (4.4) is coercive and strictly convex with respect to

$x_1, \ldots, x_T, \widetilde{x}_1, \ldots, \widetilde{x}_T$ are the unique global minimizer of the function. It follows that:

$$\sum_{t=1}^{T}(y_t - x_t)^2 + \lambda \sum_{t=p+1}^{T} |\Delta^p x_t| \geq \sum_{t=1}^{T}(y_t - \widetilde{x}_t)^2 + \lambda \sum_{t=p+1}^{T} |\Delta^p \widetilde{x}_t|, \tag{4.21}$$

where the equality holds only if $x_t = \widetilde{x}_t$ for $t = 1, \ldots, T$.[6] In addition, from (4.11) and (4.12), $y_{1-i} = \widetilde{x}_{1-i}$ for $i = 1, \ldots, g$, and $y_{T+j} = \widetilde{x}_{T+j}$ for $j = 1, \ldots, h$, we have the following inequalities:

$$\lambda \sum_{t=p+1-g}^{p+1-1} |\Delta^p x_t| \geq 0 = \lambda \sum_{t=p+1-g}^{p+1-1} |\Delta^p \widetilde{x}_t|, \tag{4.22}$$

$$\lambda \sum_{t=T+1}^{T+h} |\Delta^p x_t| \geq 0 = \lambda \sum_{t=T+1}^{T+h} |\Delta^p \widetilde{x}_t|, \tag{4.23}$$

$$\sum_{t=1-g}^{1-1} (y_t - x_t)^2 \geq 0 = \sum_{t=1-g}^{1-1} (y_t - \widetilde{x}_t)^2, \tag{4.24}$$

$$\sum_{t=T+1}^{T+h} (y_t - x_t)^2 \geq 0 = \sum_{t=T+1}^{T+h} (y_t - \widetilde{x}_t)^2. \tag{4.25}$$

Combining (4.21)–(4.23) yields

$$\sum_{t=1}^{T}(y_t - x_t)^2 + \lambda \sum_{t=p+1-g}^{T+h} |\Delta^p x_t| \geq \sum_{t=1}^{T}(y_t - \widetilde{x}_t)^2 + \lambda \sum_{t=p+1-g}^{T+h} |\Delta^p \widetilde{x}_t|, \tag{4.26}$$

where the equality in (4.26) holds only if $x_t = \widetilde{x}_t$ for $t = 1-g, \ldots, T+h$, which proves that $\widetilde{x}_t^{(\mathrm{d})} = \widetilde{x}_t$ for $t = 1-g, \ldots, T+h$. Likewise, combining (4.21)–(4.25) proves that $\widetilde{x}_t^{(\mathrm{e})} = \widetilde{x}_t$ for $t = 1-g, \ldots, T+h$ and combining (4.21) and (4.24)–(4.25) proves that $\widetilde{x}_t^{(\mathrm{f})} = \widetilde{x}_t$ for $t = 1-g, \ldots, T+h$. □

Example 4.1. *As an illustration of the above theorem, we give a numerical example. Consider the case where $T = 5$, $g = 1$, and $h = 2$. Suppose that we obtained*

$$\widetilde{x}_1 = 3, \quad \Delta\widetilde{x}_2 = 2, \quad [\Delta^2\widetilde{x}_3, \Delta^2\widetilde{x}_4, \Delta^2\widetilde{x}_5]^\top = [0, -1, 0]^\top$$

[6] In the objective function of (4.4), $\sum_{t=1}^{T}(y_t - x_t)^2$ is coercive because it is a quadratic function whose Hessian matrix is positive definite. See, e.g., Beck (2014, Lemma 2.42).

by applying ℓ_1 polynomial trend filtering of order 2 (i.e., ℓ_1 trend filtering) to a T-dimensional time-series data.[7] *Because $2 = \Delta\widetilde{x}_2 = \Delta\widetilde{x}_3 \neq \Delta\widetilde{x}_4 = \Delta\widetilde{x}_5 = 1$, the line plot of $(t, \widetilde{x}_t)$ for $t = 1, \ldots, 5$ becomes a continuous piecewise linear line such that $(3, \widetilde{x}_3)$ is a knot. $\widetilde{x}_t$ for $t = 1, \ldots, 5$ are explicitly $[\widetilde{x}_1, \widetilde{x}_2, \widetilde{x}_3, \widetilde{x}_4, \widetilde{x}_5]^\top = [3, 5, 7, 8, 9]^\top$. Then, from the above theorem, in the case, $\widetilde{x}_t^{(i)}$ for $i = \mathrm{d}, \mathrm{e}, \mathrm{f}$ and $t = 1-1, \ldots, 5+2$ are as follows:*

$$[\widetilde{x}_{1-1}^{(i)}, \widetilde{x}_1^{(i)}, \widetilde{x}_2^{(i)}, \widetilde{x}_3^{(i)}, \widetilde{x}_4^{(i)}, \widetilde{x}_5^{(i)}, \widetilde{x}_{5+1}^{(i)}, \widetilde{x}_{5+2}^{(i)}]^\top = [1, 3, 5, 7, 8, 9, 10, 11]^\top.$$

Theorem 4.2. *If $\lambda \geq 2\|(\boldsymbol{D}\boldsymbol{D}^\top)^{-1}\boldsymbol{D}\boldsymbol{y}\|_\infty$, for $i = \mathrm{d}, \mathrm{e}, \mathrm{f}$ and $t = 1-g, \ldots, T+h$, it follows that*

$$\widetilde{x}_t^{(i)} = \widehat{\beta}_0 t^0 + \cdots + \widehat{\beta}_{p-1} t^{p-1}, \tag{4.27}$$

where $(\widehat{\beta}_0, \ldots, \widehat{\beta}_{p-1}) = \arg\min_{\beta_0, \ldots, \beta_{p-1}} \sum_{t=1}^T (y_t - \beta_0 t^0 - \cdots - \beta_{p-1} t^{p-1})^2$.

Proof. Because $\boldsymbol{D}_{g+T+h}$ is a $(g+T+h-p) \times (g+T+h)$ $(p+1)$-diagonal Toeplitz matrix, such that:

$$\boldsymbol{D}_{g+T+h} = \begin{bmatrix} a_0 & \cdots & a_p & 0 & \cdots & 0 \\ 0 & \ddots & & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & & \ddots & 0 \\ 0 & \cdots & 0 & a_0 & \cdots & a_p \end{bmatrix},$$

where $a_k = (-1)^{p-k}\binom{p}{k}$ for $k = 0, \ldots, p$, it may be expressed as

$$\boldsymbol{D}_{g+T+h} = \begin{bmatrix} \boldsymbol{G}_1 & \boldsymbol{G}_2 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{D} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{H}_1 & \boldsymbol{H}_2 \end{bmatrix},$$

where $\boldsymbol{G}_1$ is a $g \times g$ upper triangular matrix, $\boldsymbol{G}_2$ is a $g \times T$ matrix, $\boldsymbol{H}_1$ is an $h \times T$ matrix, and $\boldsymbol{H}_2$

[7] In the case, $[\Delta^2\widetilde{x}_3, \Delta^2\widetilde{x}_4, \Delta^2\widetilde{x}_5]^\top$ is expected to become sparse, as in the numerical example, because $\sum_{t=3}^5 |\Delta^2 x_t|$ is included as a penalty.

is an $h \times h$ unit lower-triangular matrix. For example, when $p = 3$, $g = h = 2$, and $T = 5$:

$$\boldsymbol{D}_{2+5+2} = \left[\begin{array}{cc:ccccc:cc} -1 & 3 & -3 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 3 & -3 & 1 & 0 & 0 & 0 & 0 \\ \hdashline 0 & 0 & -1 & 3 & -3 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 3 & -3 & 1 & 0 & 0 \\ \hdashline 0 & 0 & 0 & 0 & -1 & 3 & -3 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 3 & -3 & 1 \end{array}\right]. \tag{4.28}$$

Let $\widetilde{\boldsymbol{x}}_g = [\widetilde{x}_{1-g}, \ldots, \widetilde{x}_{1-1}]^\top$, $\widetilde{\boldsymbol{x}} = [\widetilde{x}_1, \ldots, \widetilde{x}_T]^\top$, $\widetilde{\boldsymbol{x}}_h = [\widetilde{x}_{T+1}, \ldots, \widetilde{x}_{T+h}]^\top$, and $\widetilde{\boldsymbol{x}}_{g+T+h} = [\widetilde{\boldsymbol{x}}_g^\top, \widetilde{\boldsymbol{x}}^\top, \widetilde{\boldsymbol{x}}_h^\top]^\top$, which is a $(g + T + h)$-dimensional column vector. Then, by definition of $\widetilde{\boldsymbol{x}}_g$ and $\widetilde{\boldsymbol{x}}_h$, it follows that:

$$\boldsymbol{G}_1\widetilde{\boldsymbol{x}}_g + \boldsymbol{G}_2\widetilde{\boldsymbol{x}} = \boldsymbol{0}, \tag{4.29}$$

$$\boldsymbol{H}_1\widetilde{\boldsymbol{x}} + \boldsymbol{H}_2\widetilde{\boldsymbol{x}}_h = \boldsymbol{0}, \tag{4.30}$$

which leads to:

$$\boldsymbol{D}_{g+T+h}\widetilde{\boldsymbol{x}}_{g+T+h} = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{D}\widetilde{\boldsymbol{x}} \\ \boldsymbol{0} \end{bmatrix}. \tag{4.31}$$

From Kim et al. (2009), if $\lambda \geq 2\|(\boldsymbol{D}\boldsymbol{D}^\top)^{-1}\boldsymbol{D}\boldsymbol{y}\|_\infty$, it follows that $\widetilde{\boldsymbol{x}} = \boldsymbol{\Pi}\widehat{\boldsymbol{\beta}}$, where $\widehat{\boldsymbol{\beta}} = (\boldsymbol{\Pi}^\top\boldsymbol{\Pi})^{-1}\boldsymbol{\Pi}^\top\boldsymbol{y}$. Recalling that $\boldsymbol{D}\boldsymbol{\Pi} = \boldsymbol{0}$, we obtain $\boldsymbol{D}_{g+T+h}\widetilde{\boldsymbol{x}}_{g+T+h} = \boldsymbol{0}$ if $\lambda \geq 2\|(\boldsymbol{D}\boldsymbol{D}^\top)^{-1}\boldsymbol{D}\boldsymbol{y}\|_\infty$, which implies that $\widetilde{\boldsymbol{x}}_{g+T+h}$ may be represented as $\boldsymbol{\Pi}_{g+T+h}\boldsymbol{\gamma}$. Because $\widetilde{\boldsymbol{x}} = \boldsymbol{\Pi}\widehat{\boldsymbol{\beta}}$, $\boldsymbol{\gamma}$ must equal $\widehat{\boldsymbol{\beta}}$. Therefore, if $\lambda \geq 2\|(\boldsymbol{D}\boldsymbol{D}^\top)^{-1}\boldsymbol{D}\boldsymbol{y}\|_\infty$, then $\widetilde{\boldsymbol{x}}_{g+T+h} = \boldsymbol{\Pi}_{g+T+h}\widehat{\boldsymbol{\beta}}$. □

Theorem 4.3. *Suppose that $\boldsymbol{y} = \boldsymbol{\Pi}\boldsymbol{\alpha}$, where $\boldsymbol{\alpha} \neq \boldsymbol{0}$ is a p-dimensional column vector. Then, for $i = \mathrm{d}, \mathrm{e}, \mathrm{f}$, it follows that:*

$$\widetilde{\boldsymbol{x}}^{(i)}_{g+T+h} = \boldsymbol{\Pi}_{g+T+h}\boldsymbol{\alpha}, \tag{4.32}$$

where $\widetilde{\boldsymbol{x}}^{(i)}_{g+T+h} = [\widetilde{x}^{(i)}_{1-g}, \ldots, \widetilde{x}^{(i)}_{T+h}]^\top$.

Proof. If $\boldsymbol{y} = \boldsymbol{\Pi\alpha}$, it follows that: $\widetilde{\boldsymbol{x}} = \boldsymbol{\Pi\alpha}$. Accordingly, $\boldsymbol{D}_{g+T+h}\widetilde{\boldsymbol{x}}_{g+T+h} = \boldsymbol{0}$, which indicates that $\widetilde{\boldsymbol{x}}_{g+T+h}$ may be represented as $\boldsymbol{\Pi}_{g+T+h}\boldsymbol{\gamma}$. Because $\widetilde{\boldsymbol{x}} = \boldsymbol{\Pi\alpha}$ if $\boldsymbol{y} = \boldsymbol{\Pi\alpha}$, $\boldsymbol{\gamma}$ must equal $\boldsymbol{\alpha}$. Therefore, we obtain $\widetilde{\boldsymbol{x}}_{g+T+h} = \boldsymbol{\Pi}_{g+T+h}\boldsymbol{\alpha}$ if $\boldsymbol{y} = \boldsymbol{\Pi\alpha}$. □

Example 4.2. *We give a numberical example for the case* $g = 1$*,* $T = 2$*,* $h = 2$ *and* $p = 3$*. Suppose* $\boldsymbol{\alpha} = [-1, 1, 3]^\top$*, then* $\boldsymbol{y} = \boldsymbol{\Pi}_2\boldsymbol{\alpha} = [3, 13]^\top$ *is satisfied.*

From the above theorem, we obtain

$$\widetilde{\boldsymbol{x}}_{1+2+2} = \begin{bmatrix} (1-1)^0 & (1-1)^1 & (1-1)^2 \\ (1)^0 & 1^1 & 1^2 \\ (2)^0 & 2^1 & 2^2 \\ (3)^0 & 3^1 & 3^2 \\ (3+1)^0 & 4^1 & 4^2 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \\ 3 \end{bmatrix} = \begin{bmatrix} -1 \\ 3 \\ 13 \\ 29 \\ 51 \end{bmatrix}$$

Corollary 4.1. *Let* $\widetilde{\boldsymbol{x}}^{(i)}_{g+T+h} = [\widetilde{x}^{(i)}_{1-g}, \ldots, \widetilde{x}^{(i)}_{T+h}]^\top$ *for* $i = \mathrm{d}, \mathrm{e}, \mathrm{f}$.

(i) Denote the $(j+1)$*-th column of* $\boldsymbol{\Pi}$ *and that of* $\boldsymbol{\Pi}_{g+T+h}$*, respectively, by* $\boldsymbol{\tau}_j$ *and by* $\boldsymbol{\tau}_{g+T+h,j}$ *for* $j = 0, \ldots, p-1$*. If* $\boldsymbol{y} = \boldsymbol{\tau}_j$*, then* $\widetilde{\boldsymbol{x}}^{(i)}_{g+T+h} = \boldsymbol{\tau}_{g+T+h,j}$ *for any* $\lambda > 0$.

(ii) Let $\boldsymbol{z}$ *be a* T*-dimensional column vector. If* $\boldsymbol{y} = \boldsymbol{\Pi}(\boldsymbol{\Pi}^\top\boldsymbol{\Pi})^{-1}\boldsymbol{\Pi}^\top\boldsymbol{z}$*, then* $\widetilde{\boldsymbol{x}}^{(i)}_{g+T+h} = \boldsymbol{\Pi}_{g+T+h}(\boldsymbol{\Pi}^\top\boldsymbol{\Pi})^{-1}\boldsymbol{\Pi}^\top\boldsymbol{z}$ *for any* $\lambda > 0$.

4.4 Some remarks on the main results

First, we give a remark on Theorem 4.1. Because $|\boldsymbol{G}_1| = (-1)^{g\cdot p}$, from (4.29), $\widetilde{\boldsymbol{x}}_g$ may be expressed with $\widetilde{\boldsymbol{x}}$ as $\widetilde{\boldsymbol{x}}_g = -\boldsymbol{G}_1^{-1}\boldsymbol{G}_2\widetilde{\boldsymbol{x}}$. Likewise, because $|\boldsymbol{H}_2| = 1$, from (4.30), $\widetilde{\boldsymbol{x}}_h$ may be expressed with $\widetilde{\boldsymbol{x}}$ as $\widetilde{\boldsymbol{x}}_h = -\boldsymbol{H}_2^{-1}\boldsymbol{H}_1\widetilde{\boldsymbol{x}}$. Thus, the modified ℓ_1 polynomial trend filtering, (4.16), may be characterized as a filtering that calculates

$$\begin{bmatrix} -\boldsymbol{G}_1^{-1}\boldsymbol{G}_2 \\ \boldsymbol{I}_T \\ -\boldsymbol{H}_2^{-1}\boldsymbol{H}_1 \end{bmatrix} \widetilde{\boldsymbol{x}} \tag{4.33}$$

from $\boldsymbol{y}$.[8] ($\boldsymbol{I}_T$ is defined as an $T \times T$ identity matrix.)

In addition, from Kim et al. (2009), it follows that $\widetilde{\boldsymbol{x}} \to \boldsymbol{y}$ as $\lambda \to 0$. Therefore, we obtain:

$$\widetilde{\boldsymbol{x}}^{(\mathrm{d})}_{g+T+h} \to \begin{bmatrix} -\boldsymbol{G}_1^{-1}\boldsymbol{G}_2 \\ \boldsymbol{I}_T \\ -\boldsymbol{H}_2^{-1}\boldsymbol{H}_1 \end{bmatrix} \boldsymbol{y}, \quad (\lambda \to 0). \tag{4.34}$$

Second, we provide a remark on Theorems 4.2 and 4.3. Yamada (2017b) recently showed that:

$$\widetilde{\boldsymbol{x}} = \boldsymbol{\Pi}\widehat{\boldsymbol{\beta}} + \boldsymbol{X}\widetilde{\boldsymbol{\phi}}, \tag{4.35}$$

where $\boldsymbol{X} = \boldsymbol{D}^\top(\boldsymbol{D}\boldsymbol{D}^\top)^{-1}$ and $\widetilde{\boldsymbol{\phi}}$, which is a $(T-p)$-dimensional column vector, is the solution of the following Lasso regression/BPDN:

$$\min_{\boldsymbol{\phi}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\phi}\|_2^2 + \lambda\|\boldsymbol{\phi}\|_1. \tag{4.36}$$

Because $\boldsymbol{X}^\top\boldsymbol{\Pi} = \boldsymbol{0}$, $\boldsymbol{\Pi}\widehat{\boldsymbol{\beta}} + \boldsymbol{X}\widetilde{\boldsymbol{\phi}}$ in (4.35) represents an orthogonal decomposition of $\widetilde{\boldsymbol{x}}$. Here, we show that we may prove Theorems 4.2 and 4.3 by using (4.35) and (4.36). Premultiplying (4.35) by $\boldsymbol{D}$ yields $\boldsymbol{D}\widetilde{\boldsymbol{x}} = \widetilde{\boldsymbol{\phi}}$. We accordingly obtain:

$$\boldsymbol{D}_{g+T+h}\widetilde{\boldsymbol{x}}_{g+T+h} = \begin{bmatrix} \boldsymbol{0} \\ \widetilde{\boldsymbol{\phi}} \\ \boldsymbol{0} \end{bmatrix}. \tag{4.37}$$

(i) From Osborne et al. (2000, p. 324), if $\lambda \geq 2\|\boldsymbol{X}^\top\boldsymbol{y}\|_\infty$, then $\widetilde{\boldsymbol{\phi}} = \boldsymbol{0}$. Therefore, we obtain $\widetilde{\boldsymbol{x}} = \boldsymbol{\Pi}\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{D}_{g+T+h}\widetilde{\boldsymbol{x}}_{g+T+h} = \boldsymbol{0}$, which proves Theorem 4.2.

[8] Let us calculate $-\boldsymbol{H}_2^{-1}\boldsymbol{H}_1\widetilde{\boldsymbol{x}}$ for the case where $p=3$, $g=h=2$, and $T=5$. From (4.28), it follows that

$$-\boldsymbol{H}_1\widetilde{\boldsymbol{x}} = \begin{bmatrix} \widetilde{x}_{T-2} - 3\widetilde{x}_{T-1} + 3\widetilde{x}_T \\ \widetilde{x}_{T-1} - 3\widetilde{x}_T \end{bmatrix} = \begin{bmatrix} \widetilde{x}_T + (\Delta\widetilde{x}_T) + (\Delta^2\widetilde{x}_T) \\ -2\widetilde{x}_T - (\Delta\widetilde{x}_T) \end{bmatrix}.$$

Accordingly, we obtain:

$$-\boldsymbol{H}_2^{-1}\boldsymbol{H}_1\widetilde{\boldsymbol{x}} = \begin{bmatrix} 1 & 0 \\ 3 & 1 \end{bmatrix} \begin{bmatrix} \widetilde{x}_T + (\Delta\widetilde{x}_T) + (\Delta^2\widetilde{x}_T) \\ -2\widetilde{x}_T - (\Delta\widetilde{x}_T) \end{bmatrix} = \begin{bmatrix} \widetilde{x}_T + (\Delta\widetilde{x}_T) + (\Delta^2\widetilde{x}_T) \\ \widetilde{x}_T + 2(\Delta\widetilde{x}_T) + 3(\Delta^2\widetilde{x}_T) \end{bmatrix},$$

which is consistent with (4.15).

(ii) If $\boldsymbol{y} = \boldsymbol{\Pi}\boldsymbol{\alpha}$, where $\boldsymbol{\alpha} \neq \boldsymbol{0}$, then $\boldsymbol{X}^\top \boldsymbol{y} = \boldsymbol{0}$, which implies that $\lambda > 2\|\boldsymbol{X}^\top \boldsymbol{y}\|_\infty = 0$. Again, from Osborne et al. (2000), we obtain $\widetilde{\boldsymbol{\phi}} = \boldsymbol{0}$ if $\boldsymbol{y} = \boldsymbol{\Pi}\boldsymbol{\alpha}$. Therefore, if $\boldsymbol{y} = \boldsymbol{\Pi}\boldsymbol{\alpha}$, it follows that $\widetilde{\boldsymbol{x}} = \boldsymbol{\Pi}\widehat{\boldsymbol{\beta}} = \boldsymbol{\Pi}\boldsymbol{\alpha}$ and $\boldsymbol{D}_{g+T+h}\widetilde{\boldsymbol{x}}_{g+T+h} = \boldsymbol{0}$, which proves Theorem 4.3.

Example 4.3. *Third, we give an example of Corollary 4.1 (i). For the case where $\boldsymbol{y} = [1, \ldots, 5]^\top$ and $p = g = h = 2$, it follows that $\widetilde{\boldsymbol{x}}^{(\mathrm{d})}_{2+5+2} = [-1, 0, 1, \ldots, 5, 6, 7]^\top$ for any $\lambda > 0$.*

4.5 Concluding remarks

The ℓ_1 polynomial trend filtering method is a promising piecewise polynomial curve-fitting method because it does not require prespecifying the number and location of knots. We have shown some theoretical results on this method. One of them is that a small modification of the filtering provides identical trend estimates and also extrapolations of the trend beyond both sample limits. Another is that $\widetilde{x}_{T+1}, \ldots, \widetilde{x}_{T+h}$ based on (4.12) are useless to improve the trend estimates of ℓ_1 polynomial trend filtering. We also provided a MATLAB function for calculating the solution of one of the modified filtering methods. The main results of the paper are summarized in Theorems 4.1, 4.2, and 4.3 and Corollary 4.1.

Finally, we remark that applying the modified ℓ_1 polynomial trend filtering (4.16)–(4.18) requires specification of the value of λ. For this purpose, the methods proposed in Yamada and Yoon (2016b) and Yamada (2018) are applicable.

4.6 Appendix

4.6.1 Proof of (4.15)

Because $\Delta^3\widetilde{x}_{T+j} = \Delta^2\widetilde{x}_{T+j} - \Delta^2\widetilde{x}_{T+j-1}$, from $\Delta^3\widetilde{x}_{T+j} = 0$ for $j = 1, \ldots, h$, we obtain $\Delta^2\widetilde{x}_{T+k} = \Delta^2\widetilde{x}_T$ for $k = 1, \ldots, h$. Then, because $\sum_{k=1}^{l}(\Delta^2\widetilde{x}_{T+k}) = l(\Delta^2\widetilde{x}_T)$ for $l = 1, \ldots, h$ and $\sum_{k=1}^{l}(\Delta^2\widetilde{x}_{T+k}) = \Delta\widetilde{x}_{T+l} - \Delta\widetilde{x}_T$, it follows that

$$\Delta\widetilde{x}_{T+l} = \Delta\widetilde{x}_T + l(\Delta^2\widetilde{x}_T), \quad (l = 1, \ldots, h).$$

Furthermore, because $\sum_{l=1}^{j}(\Delta\widetilde{x}_{T+l}) = j(\Delta\widetilde{x}_T) + (\sum_{l=1}^{j} l)(\Delta^2\widetilde{x}_T)$ for $j = 1, \ldots, h$ and $\sum_{l=1}^{j}(\Delta\widetilde{x}_{T+l}) = \widetilde{x}_{T+j} - \widetilde{x}_T$, we finally obtain:

$$\widetilde{x}_{T+j} = \widetilde{x}_T + j(\Delta\widetilde{x}_T) + \frac{j(j+1)}{2}(\Delta^2\widetilde{x}_T), \quad (j = 1, \ldots, h).$$

4.7 References

1. Beck, A., 2014, *Introduction to nonlinear optimization theory, algorithms, and applications with MATLAB*, SIAM, Philadelphia.

2. Chen, S. S., D. L. Donoho, and M. A. Saunders, 1998, Atomic decomposition by basis pursuit, *SIAM Journal on Scientific Computing*, 20, 1, 33–61.

3. Grant, M. and S. Boyd, 2013, CVX: Matlab Software for Disciplined Convex Programming, version 2.0 beta.
   URL:`http://cvxr.com/cvx`.

4. Harchaoui, Z. and C. Lévy-Leduc, 2010, Multiple change-point estimation with a total variation penalty, *Journal of the American Statistical Association*, 105, 492, 1480–1493.

5. Hodrick, R. J. and E. C. Prescott, 1997, Postwar U.S. business cycles: An empirical investigation, *Journal of Money, Credit and Banking*, 29, 1, 1–16.

6. Hoerl, A. E. and R. W. Kennard, 1970, Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics*, 12, 1, 55–67.

7. Kim, S., K. Koh, S. Boyd, and D. Gorinevsky, 2009, ℓ_1 trend filtering, *SIAM Review*, 51, 2, 339–360.

8. Koenker, R., P. Ng, and S. Portnoy, 1994, Quantile smoothing splines, *Biometrika*, 81, 4, 673–680.

9. Miller, M. D., 1946, *Elements of graduation*, Actuarial Society of America and American Institute of Actuaries, Philadelphia.

10. Mohr, M., 2005, A trend-cycle(-season) filter, European Central Bank Working Paper, No. 499.

11. Nocon, A. S. and W. F. Scott, 2012, An extension of the Whittaker–Henderson method of graduation, *Scandinavian Actuarial Journal*, 2012, 1, 70–79.

12. Osborne, M. R., B. Presnell, and B. A. Turlach, 2000, On the LASSO and its dual, *Journal of Computation and Graphical Statistics*, 9, 2, 319–337.

13. Phillips, P. C. B., 2010, Two New Zealand pioneer econometricians, *New Zealand Economic Papers*, 44, 1, 1–26.

14. Schuette, D. R., 1978, A linear programming approach to graduation, *Transactions of Society of Actuaries*, 30, 407–431.

15. Tibshirani, R., 1996, Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society Series B*, 58, 1, 267–288.

16. Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and K. Knight, 2005, Sparsity and smoothness via the fused lasso, *Journal of the Royal Statistics Society: Series B*, 67, 1, 91–108.

17. Tibshirani, R. J. and J. Taylor, 2011, The solution path of the generalized lasso, *Annals of Statistics*, 39, 3, 1335–1371.

18. Tibshirani, R. J., 2014, Adaptive piecewise polynomial estimation via trend filtering, *Annals of Statistics*, 42, 1, 285–323.

19. Weinert, H. L., 2007, Efficient computation for Whittaker–Henderson smoothing, *Computational Statistics and Data Analysis*, 52, 2, 959–974.

20. Winkelried, D., 2016, Piecewise linear trends and cycles in primary commodity prices, *Journal of International Money and Finance*, 64, 196–213.

21. Yamada, H., 2015, Ridge regression representations of the generalized Hodrick–Prescott filter, *Journal of the Japan Statistical Society*, 45, 2, 121–128.

22. Yamada, H., 2017a, Estimating the trend in US real GDP using the ℓ_1 trend filtering, *Applied Economics Letters*, 24, 10, 713–716.

23. Yamada, H., 2017b, A trend filtering method closely related to ℓ_1 trend filtering, *Empirical Economics*, 55, 4, 1413–1423.

24. Yamada, H., 2017c, A small but practically useful modification to the Hodrick–Prescott filtering: A note, *Communications in Statistics – Theory and Methods*, 46, 17, 8430–8434.

25. Yamada, H., 2018, A new method for specifying the tuning parameter of ℓ_1 trend filtering, *Studies in Nonlinear Dynamics and Econometrics*, 22, 4, 1–8.

26. Yamada, H. and R. Du, 2018, Some results on ℓ_1 polynomial trend filtering, *Econometrics*, 6, 3, 33.

27. Yamada, H. and L. Jin, 2013, Japan's output gap estimation and ℓ_1 trend filtering, *Empirical Economics*, 45, 1, 81–88.

28. Yamada, H. and G. Yoon, 2014, When Grilli and Yang meet Prebisch and Singer: Piecewise linear trends in primary commodity prices, *Journal of International Money and Finance*, 42, 193–207.

29. Yamada, H. and G. Yoon, 2016a, Measuring the US NAIRU as a step function, *Empirical Economics*, 51, 4, 1679–1688.

30. Yamada, H. and G. Yoon, 2016b, Selecting the tuning parameter of the ℓ_1 trend filter, *Studies in Nonlinear Dynamics and Econometrics*, 20, 1, 97–105.