Model selection criteria for regression models with splines and the automatic localization of knots

Alex Rodrigo dos S. Sousa   Magno T. F. Severino   Florencia Leonardi

June 3, 2020

Abstract

In this paper we propose a model selection approach to fit a regression model using splines with a variable number of knots. We introduce a penalized criterion to estimate the number and the position of the knots where to anchor the splines basis. The method is evaluated on simulated data and applied to covid-19 daily reported cases for short-term prediction.

1 Introduction

Spline regression is one of the most used tools for nonparametric and semiparametric regression and has been extensively developed along the last decades. The main ideia is to select some knots on the function domain to define a partition and then fit a fixed degree polynomial on each segment in such a way that these polynomials smoothly match at the internal knots through continuity conditions imposed on their derivatives. The result is typically a smooth curve fitting, which accurately estimates a large variety of smooth functions. Further, spline regression offers a great flexibility in curve estimation problems according to convenient choices of degree and internal knots and, as generally occurs in nonparametric regression by basis expansion, the infinite estimation problem becomes a finite parametric estimation problem, once the spline function is determined by the estimation of its coefficients. For more details about spline based methods, their theory and applications in Statistics, see Wahba (1990), Ramsay (2004), Friedman et al. (2001), and James et al. (2013).

One of the main issues in spline regression is the selection of the number and positions of the knots. In some data analyses, wrong specification of the knots can cause a high impact on the spline regression performance. Excessive number of knots may lead the spline regression to overfit the dataset while a not enough number of knots can lead to underfitting. Moreover, the positions of the knots should be well determined, since high concentration of knots on a specific interval of the domain overfits the data on this region and underfits the data outside it. Then, a good procedure for knots selection should provide optimal number and locations of knots, in order to have a well performed spline curve fitting.

In practice, the selection of knots is done by hand and few studies in the literature provide some systematic method to overcome this difficulty. Friedman and Silverman (1989), Friedman (1991) and Stone et al. (1997) studied stepwise based methods for a set of candidates of knots. Osborne et al. (1998) proposed an algorithm based on the LASSO estimator to estimate the locations of knots. Ruppert (2002) presented two algorithms for selection of a fixed number of knots on penalized spline regression, called P-splines. The first one is based on the minimization of the generalized cross validation statistic and the second one is based on the Demmler-Reinsch type diagonalization and Ülker and Arslan (2009) proposed an artificial immune system to perform knots selection in curve and surface B-splines estimation by considering the knots locations candidates as antibodies. Bayesian methods are proposed by Denison et al. (1998) who proposed a joint prior distribution for the number and locations of knots, and Biller (2000), who gave an adaptive bayesian approach for knot selection based on reversible jump Markov chain Monte Carlo in semiparametric generalized linear models.

Although the methods described above have been well succeeded, they work with a fixed number of knots and/or require some prior information regarding possible regions where the knots should be set, which can be seen as a limitation in some cases. In this work we propose an automatic method to estimate both the number of knots and their locations, based on the minimization of a penalized least squares criterion. The penalty plays the role of avoiding a high number of knots and consequently to overfit the data. We test the performance of the method on simulated data and we also apply it to daily reported cases of covid-19 on several countries to obtain a fit on the epidemiological curve.

This paper is organized as follows: Section 2 is addressed to the description of the spline model and the proposed method of automatic knots selection. Simulation studies about the performance of the method are shown in Section 3. Applications to real datasets involving daily reported cases of Covid-19 are done in Section 4. Section 5 provides conclusion and suggestions of future works.

2 Spline regression model with variable number of knots

We start with a nonparametric regression problem of the form

$$y_i = f(x_i) + \epsilon_i, \qquad i = 1, ..., n,$$

where x_i are scalars, f is an unknown smooth function with domain $[a, b]$ such that $f(x_i) = \mathbb{E}(y_i|x_i)$ and $\{\epsilon_i\}_{i=1}^n$ are zero mean independent random variables with unknown variance σ^2. To estimate f, we define a sequence of knots $a = t_0 < t_1 < ... < t_K < t_{K+1} = b$ that defines a partition of $[a, b]$ in intervals $[a, t_1), [t_1, t_2), ..., [t_K, b]$ and use a $m-$th order spline regression model, $m = p + 1$, defined by

$$f(x) = \beta_0 + \sum_{j=1}^{p} \beta_j x^j + \sum_{k=1}^{K} \beta_{p+k}(x - t_k)_+^p, \qquad (2.1)$$

where $p, K \in \mathbb{Z}_+$, $\boldsymbol{\beta} = [\beta_0, ..., \beta_{p+K}]'$ is the vector of coefficients, $(u)_+^p = u^p \mathbb{I}(u \geq 0)$ and $t_1 < ... < t_K$ are fixed internal knots which, from the now on, will be referred to as just knots. In general, a spline regression

model is a p degree polynomial on each interval of two consecutive knots and has $p-1$ continuous derivatives everywhere (supposing that the knots have no multiplicity). In this work we adopt a set of knots sufficiently spaced, i.e, we assume $|t_{k+1} - t_k| > \delta$ for some $\delta > 0$, $k = 0, ..., K$. Thus, the problem of estimating f becomes a finite parametric estimation problem, whose parameters are the vector of coefficients $\boldsymbol{\beta}$, the number of interna knots K and their locations t_k, $k = 1, ..., K$. Let us denote by $\boldsymbol{\theta}$ the vector of parameters to be estimated, i.e, $\boldsymbol{\theta} = [\boldsymbol{\beta}, K, t_1, ..., t_K]'$. We assume a fixed value for p, the degree of the spline function.

The estimation of the parameters should consider its performance regarding a loss function, such as in our case, the the residual sum of squares and also some degree of smoothness for f. We propose to control the smoothness with a roughness penalty defined in terms of the number of intervals established by the knots, i.e, the K knots define $K + 1$ intervals on the domain of f. Thus, we define the penalized sum of squares PSS_λ as

$$PSS_\lambda(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{i=1}^{n}[y_i - f(x_i)]^2 + \lambda(K + 1), \tag{2.2}$$

where $\boldsymbol{y} = [y_1, ..., y_n]'$, $\lambda > 0$ is a tuning parameter previously chosen and f is the spline function (2.1). Thus, the estimator $\hat{\boldsymbol{\theta}}$ is obtained by

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} PSS_\lambda(\boldsymbol{\theta}; \boldsymbol{y}). \tag{2.3}$$

Considering the definition above, we search for a curve estimate that has good fitting properties with low residual sum of squares and which is also smooth enough according to a reasonable number of knots that is controlled by the penalty on the number of segments of the domain. The tuning parameter λ attributes a weight on the roughness penalty term. High λ values increase the importance of the penalization, i.e, a preference for smoothness is given against accuracy. On the other hand, low λ value favors models with high number of parameters, which can lead to overfitting. The specification of a good value for λ is an issue inherent to any regularized estimator and, as in other similar approaches, could be chosen by a standard cross validation procedure.

2.1   B-splines and natural splines models

Our proposed method for automatic knots selection can be applied to other spline basis structures. A well known spline basis is the B-splines, proposed by De Boor et al. (1978), which are more adapted for computational implementation due to its compact support. The i-th B-spline of order m and knots $a = t_0 < t_1 < ... < t_K < t_{K+1} = b$, denoted by $B_{i,m}(x)$ is a spline of order m defined on the same knots such that is nonzero over at most m consecutive subintervals, i.e, over $[t_i, t_{i+m}]$ and zero outside it and can be defined recursively by

$$B_{i,m}(x) = \frac{x - t_i}{t_{i+m} - t_i}B_{i,m-1}(x) + \frac{t_{i+m+1} - x}{t_{i+m+1} - t_{i+1}}B_{i+1,m-1}(x), \tag{2.4}$$

where $B_{i,1}(x) = \mathbb{I}_{[t_i, t_{i+1})}(x)$. Then, f is expanded in the B-splines basis instead of (2.1). Figure 2.1 presents eight cubic B-splines (B-splines of order 4) basis functions defined by four interior knots on the interval $[0, 10]$.
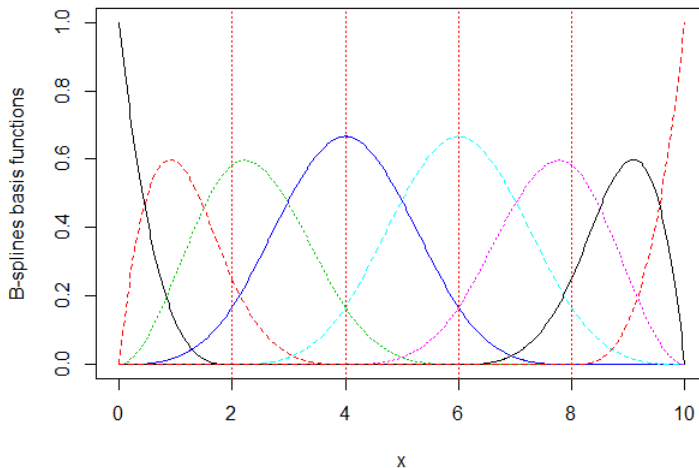
Figure 2.1: Eight cubic B-splines basis defined by four interior knots.

Another commonly used spline basis structure is the natural spline. This approach imposes the function f to be linear on its extremes, i.e, in regions lesser than t_1 and greater than t_K. Such constraint guarantees stable estimation on the boundaries of the function, see James et al. (2013). For more about spline basis structures, see De Boor (2001) .

To illustrate the role of the penalization on the curve estimation, we generate data from an underlying function defined by cubic B-splines basis on $[0, 100]$ with three not equally spaced knots at $t_1 = 20$, $t_2 = 45$, and $t_3 = 80$ and signal to noise ratio SNR=3. Figure 2.2(a) shows the underlying function, its knots locations and the generated data. After data generation, we estimate the underlying function using cubic B-splines expansion considering eleven knots with prespecified locations fixed at $[t_1, t_2, \cdots, t_{11}]' = [6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26]'$. Note that we intentionally set several knots at one specific region of domain, around the higher peak of the curve. The estimated curve is the pink line in the Figure 2.2(b). We can observe the impact of two possible issues related with bad number of knots determination and their locations. First, the estimated curve overfits data at the region around the higher peak, the location where we set the most knots, i.e, the curve fit did not recover the peak with the required degree of smoothness and provided some kind of interpolation there. On the other hand, the absence of knots located at the remaining domain did not allow the estimated curve to identify the features of the underlying function on that region, i.e, the estimated curve is too smooth to identify peaks and valleys of the function on the subinterval of the domain with no knots positions, which characterizes an underfit phenomenon. For this reason, it is important to consider a penalty term that takes account this trade off between number/locations of knots and smoothness. Penalizing the least squares by the number of segments determined by the knots partition guarantees the estimated curve to have this optimal number/locations of knots to avoid both under and overfitting prob-

lems. Moreover, this procedure is automatically performed, thus requiring no previous knowledge regarding the number and positions of knots, which is of great interest and an advantage of our method. The blue line of Figure 2.2(b) is the estimated curve provided by the proposed method. The number of knots was correctly determined, $\hat{K} = 3$ and the estimated knots locations were $\hat{t}_1 = 21$, $\hat{t}_2 = 47$ and $\hat{t}_3 = 79$.
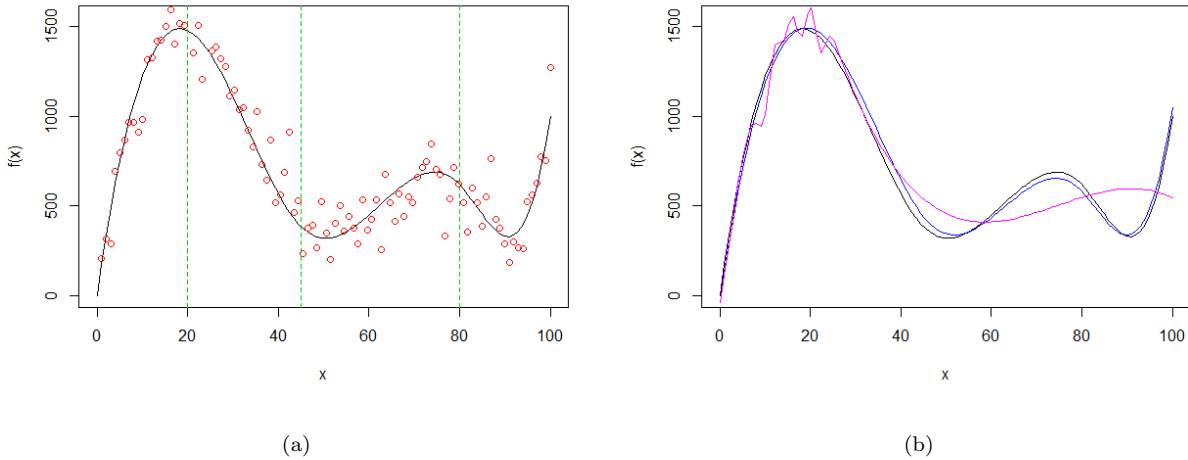

(a)	(b)

Figure 2.2: (a) Underlying function (black line), data points generated by this function and knots locations (green vertical dashed lines). (b) Underlying function (black), estimated curve by cubic B-splines with 11 knots at $[t_1, t_2, \cdots, t_{11}]' = [6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26]'$ (pink) characterizing locally overfitting around the higher peak and undefitting on remaining domain and estimated curve by the proposed method (blue) with optimal number and locations of knots.

3   Simulation studies

The proposed method was applied in simulated data to evaluate its performance in several scenarios. We considered the cubic B-splines with one knot at $t_1 = 50$, two knots at $t_1 = 25$ and $t_2 = 75$ and three knots at $t_1 = 25$, $t_2 = 50$ and $t_3 = 75$ as underlying functions. These functions are shown in Figure 3.1.

(a) $t_1 = 50$.   (b) $t_1 = 25$ and $t_2 = 75$.   (c) $t_1 = 25$, $t_2 = 50$ and $t_3 = 75$.
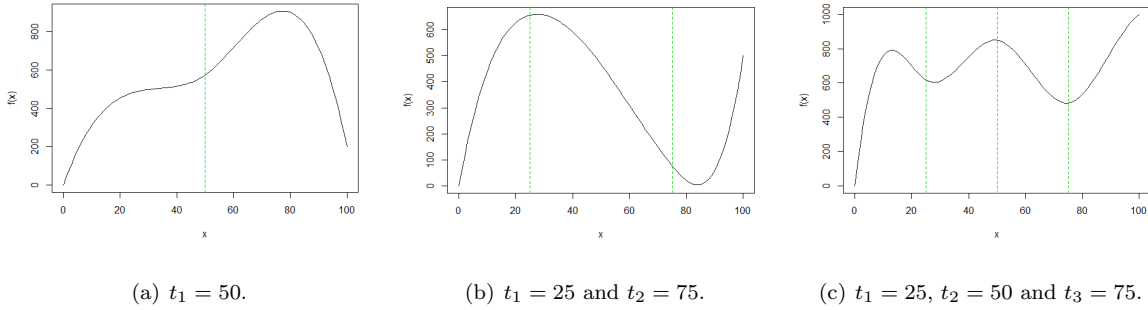
Figure 3.1: Underlying functions considered in the simulation studies. The green dashed vertical lines indicate the knots locations.

From each of the functions, data were generated with the addition of zero mean normal noise with variance according to three signal to noise ratio values (SNR), SNR $= 3$, 6, and 9 and two sample sizes, $n = 100$, and 1000. After data generation, our proposed automatic knots selection method was applied to estimate the number of knots and their locations with $\delta = 15$ for minimum knots distance. This process was repeated 1000 times for each scenario and the proportion of right estimated number of knots, the sample mean, median, standard deviation (SD) for the estimated knots and a 95% confidence interval (CI) for the knots were obtained. The results are shown in Tables 3.1, 3.2, and 3.3 for the underlying function with one knot at $t_1 = 50$, two knots at $t_1 = 25$ and $t_2 = 75$ and three knots at $t_1 = 25$, $t_2 = 50$ and $t_3 = 75$ respectively.

In fact, the proposed method had a great performance in the simulations, with accurate estimation of the number and locations of the knots in practically all the scenarios considered. Even for a low SNR, in the case SNR=3, the method produced knots estimates very close to the true knots and with low standard deviation values. As expected, estimates improved as SNR and/or n increased. The curve estimates with the true values of the knots, mean and median values for the estimated knots are shown in Figure 3.2, boxplots and histograms of the estimated knots are presented in Figures 3.3 and 3.4, respectively, for $n = 100$. The behavior of the estimates for $n = 1000$ was similar.

6

n	SNR	% of $\hat{K} = 1$	Knot	Mean	Median	SD	CI(95%)
100	3	77%	1	50.18	50.00	3.19	(44.00 ; 56.00)
	6	81%	1	49.99	50.00	1.61	(47.00 ; 53.00)
	9	99%	1	49.92	50.00	1.13	(48.00 ; 52.00)
1000	3	100%	1	49.93	50.00	1.23	(48.00 ; 52.00)
	6	100%	1	50.05	50.00	0.63	(49.00 ; 51.00)
	9	100%	1	49.99	50.00	0.48	(49.00 ; 51.00)

Table 3.1: Simulation results of the proposed method for data generated from a cubic B-splines with one knot at $t_1 = 50$.

n	SNR	% of $\hat{K} = 2$	Knot	Mean	Median	SD	CI(95%)
100	3	73%	1	28.90	25.00	9.63	(21.00 ; 56.00)
			2	67.11	71.00	12.32	(42.00 ; 79.00)
	6	84%	1	24.62	23.00	4.81	(21.00 ; 37.65)
			2	67.45	71.00	12.01	(42.35 ; 79.00)
	9	91%	1	23.63	23.00	3.16	(21.00 ; 31.00)
			2	68.68	74.00	11.55	(44.00 ; 79.00)
1000	3	73%	1	27.95	26.00	7.55	(21.00 ; 46.43)
			2	74.53	75.00	2.78	(68.00 ; 79.00)
	6	86%	1	25.73	25.00	3.91	(21.00 ; 35.00)
			2	74.80	75.00	1.27	(72.00 ; 77.00)
	9	95%	1	25.17	25.00	2.55	(21.00 ; 30.00)
			2	74.93	75.00	0.92	(73.00 ; 77.00)

Table 3.2: Simulation results of the proposed method for data generated from a cubic B-splines with two knots at $t_1 = 25$ and $t_2 = 75$.

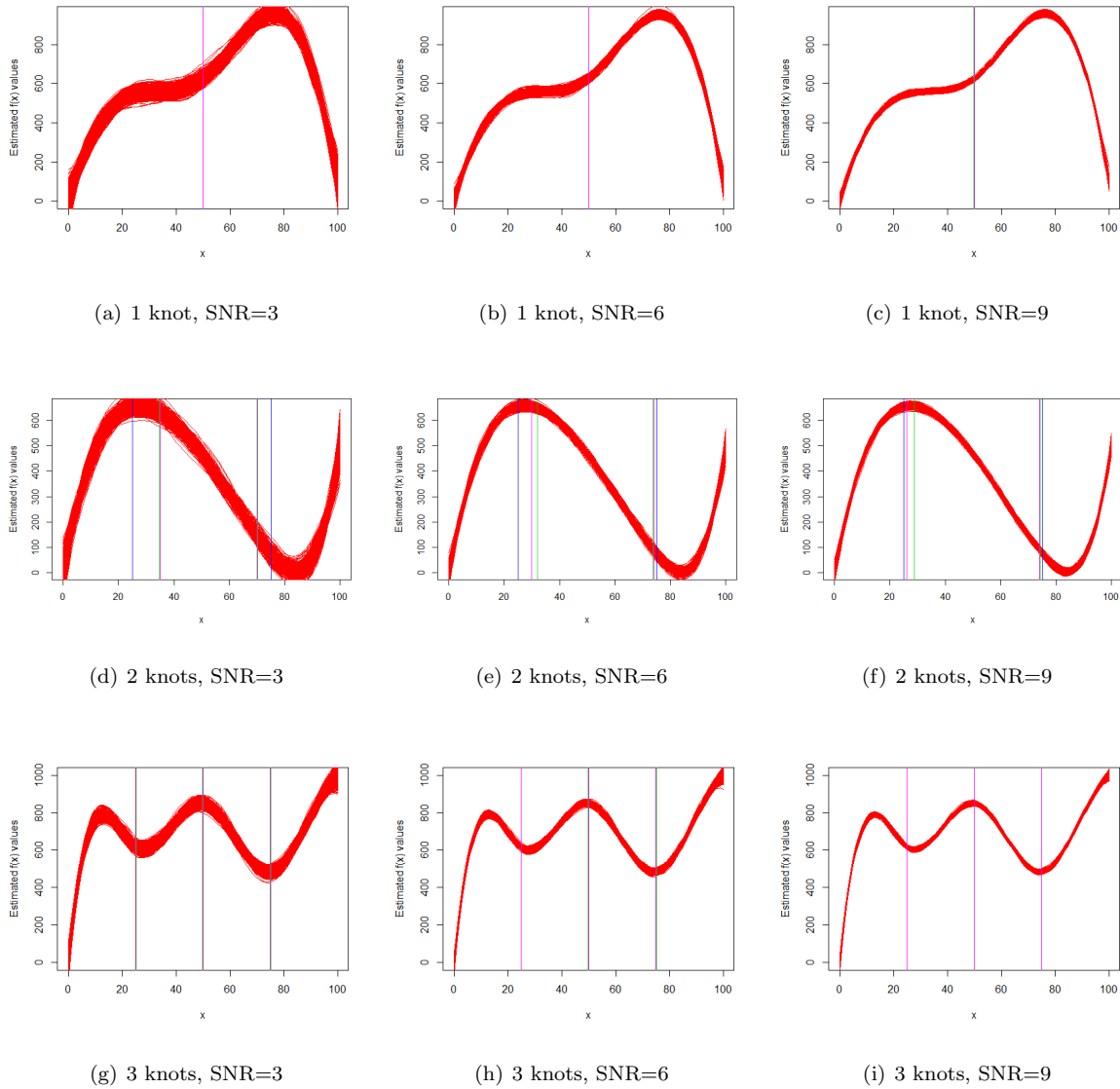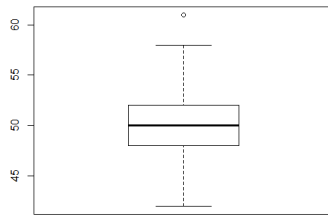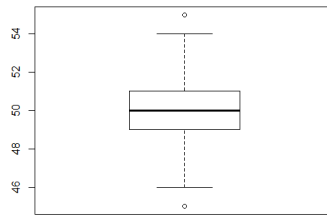| n	SNR	% of $\hat{K} = 3$	Knot	Mean	Median	SD	CI(95%)
100	3	100%	1	25.11	25.00	1.67	(22.00 ; 28.00)
			2	49.94	50.00	2.06	(46.00 ; 54.00)
			3	75.07	75.00	2.60	(70.00 ; 79.00)
	6	100%	1	25.08	25.00	0.96	(23.00 ; 27.00)
			2	49.91	50.00	1.39	(47.00 ; 52.00)
			3	75.14	75.00	1.83	(72.00 ; 79.00)
	9	100%	1	25.03	25.00	0.71	(24.00 ; 26.00)
			2	50.04	50.00	1.09	(48.00 ; 52.00)
			3	74.96	75.00	1.31	(73.00 ; 77.00)
1000	3	100%	1	25.07	25.00	0.67	(24.00 ; 26.00)
			2	49.92	50.00	1.08	(48.00 ; 52.00)
			3	75.07	75.00	1.26	(73.00 ; 77.00)
	6	100%	1	25.02	25.00	0.27	(24.48 ; 26.00)
			2	49.97	50.00	0.49	(49.00 ; 51.00)
			3	75.08	75.00	0.66	(74.00 ; 76.00)
	9	100%	1	25.00	25.00	0.10	(24.80 ; 25.30)
			2	50.01	50.00	0.23	(50.00 ; 51.00)
			3	74.99	75.00	0.34	(74.00 ; 76.00)

Table 3.3: Simulation results of the proposed method for data generated from a cubic B-splines with three knots at $t_1 = 25$, $t_2 = 50$ and $t_3 = 75$.

(a) 1 knot, SNR=3   (b) 1 knot, SNR=6   (c) 1 knot, SNR=9

(d) 2 knots, SNR=3   (e) 2 knots, SNR=6   (f) 2 knots, SNR=9

(g) 3 knots, SNR=3   (h) 3 knots, SNR=6   (i) 3 knots, SNR=9

Figure 3.2: Estimated curves for $n = 100$. Blue vertical lines are the true knots locations. Green and pink vertical lines are the mean and median of the estimated knots locations respectively. Note that in most of the scenarios, the vertical lines are overlapping.
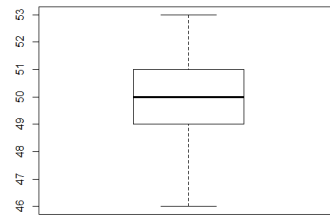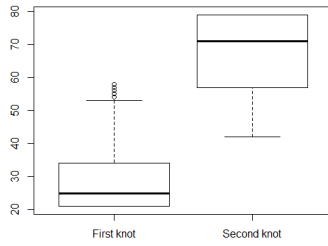
9

(a) 1 knot, SNR=3

(b) 1 knot, SNR=6

(c) 1 knot, SNR=9
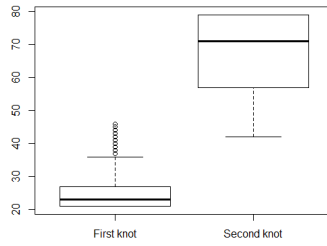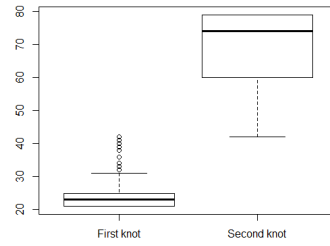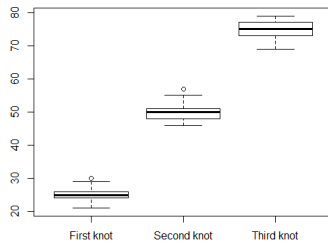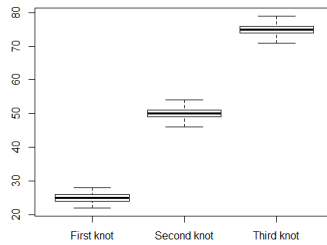
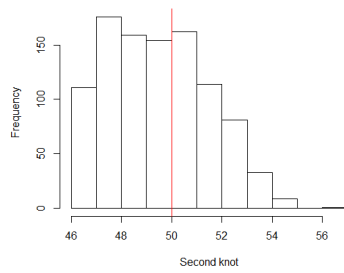(d) 2 knots, SNR=3
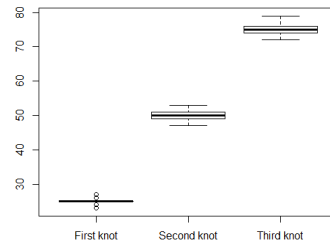
(e) 2 knots, SNR=6

(f) 2 knots, SNR=9

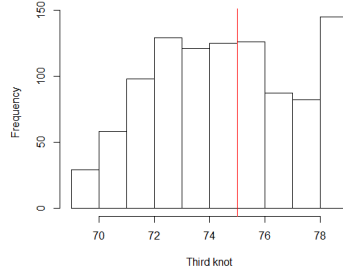(g) 3 knots, SNR=3

(h) 3 knots, SNR=6

(i) 3 knots, SNR=9
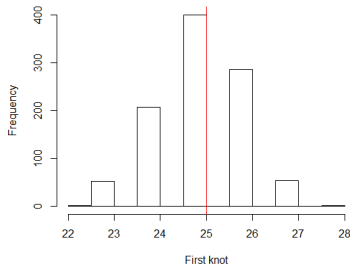
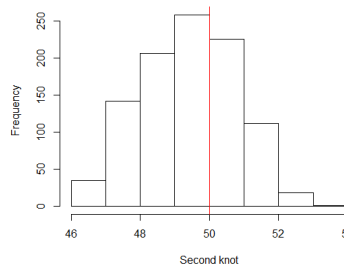Figure 3.3: Boxplots of the estimated knots for $n = 100$.
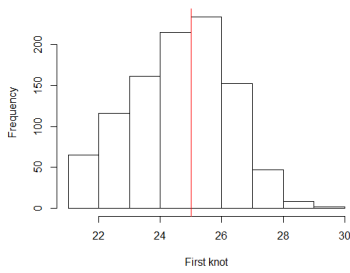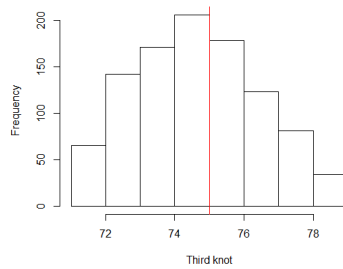
(a) SNR=3, 1st knot.
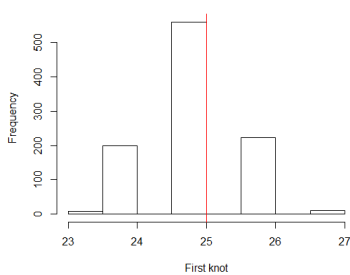
(b) SNR=3, 2nd knot.

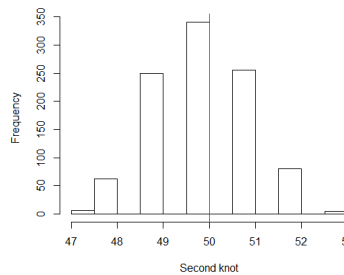(c) SNR=3, 3rd knot.

(d) SNR=6, 1st knot.
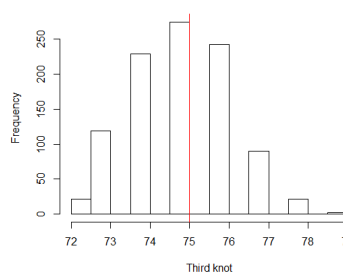
(e) SNR=6, 2nd knot.

(f) SNR=6, 3rd knot.

(g) SNR=9, 1st knot.

(h) SNR=9, 2nd knot.

(i) SNR=9, 3rd knot.

Figure 3.4: Histograms of the estimated knots for $n = 100$ and $K = 3$ knots.

4 Analysis of covid-19 time series

The counts of daily new cases of the novel coronavirus disease (covid-19) varies considerably throughout the affected countries and regions, mainly due to testing criteria used, delays in notification, and population and authorities response to the outbreak, among others. Even though, the analysis of the disease's daily new infected cases is important to understand at what stage of the epidemic each country is in. Therefore, we can use the method presented above to estimate the dates at which the trend of new cases counts had changed. This may be helpful not only to define public policy such as when safely relax social distance measures but

11

also to verify the effectiveness of mitigations measures taken. Here we perform estimation and prediction for Covid-19 data using our proposed method, the data is publicly available at `https://opendata.ecdc.europa.eu/covid19/`.

In general, during an epidemic the daily number of new infected cases shows several trends that are not related to the disease spreading itself, but to other causes, such as delay in test processing time or lack of testing. For this reason, we considered the 7-days moving average smoother on the time series before applying our method. Further, to avoid knots estimation in the beginning of the time series, we remove the 10% initial dates as possible knots locations.

Considering data available until May 30th, Figure 4.1 shows the curve fitting using natural cubic splines to data in linear scale for Brazil, USA, Spain, South Korea, Iran, Switzerland, Germany, and Italy, eight countries that are in different stages of the epidemic. Although linear scale is better for visualization of peaks and/or oscillations and natural cubic splines are suitable for a smooth fitting, the change points are better determined when we consider linear splines as splines basis to the data transformed to logarithmic scale. For this reason, we also present curve fitting for the same eight countries using linear splines on data in logarithmic scale in Figure 4.2.

In the European countries considered here, we note that mean daily new cases present a decreasing pattern as the peak occurred in late March, period in which our method selects some knots, indicating a change in trend. On the other hand, the daily new reported cases are in a high plateau in the USA, since the right-hand side of the curve is nearly parallel to the horizontal axis of Figure 4.2(b). Moreover, the first change point in late March indicates a decrease in the slope of the fitted curve, the remaining knots are located near the date the country reaches its peak so far. In Brazil, the data suggest the country have not reached its peak of infection yet, change points merely reveal small changes in the increasing pattern. Finally, Iran and South Korea both show oscillations during the observed period, however at latest dates the new cases are apparently increasing, suggesting a new wave of reported cases. We also perform similar analysis for several other countries and for Brazilian states and municipalities, that are updated daily, and the results and coded algorithms in R can be accessed at the webpage `https://www.ime.usp.br/~gpeca/covid-19`.

We mention that the principal aim of this section is solely to show a descriptive analysis of the Covid-19 situation in the countries based on public available datasets and the proposed automatic method of knots estimation as a tool to do so. For a public policy guide to deal with this sad pandemic context, several social and biological factors beyond the scope of our study should be considered on a deeper analysis.

(a) Brazil.

(b) USA.

(c) Spain.

(d) South Korea.
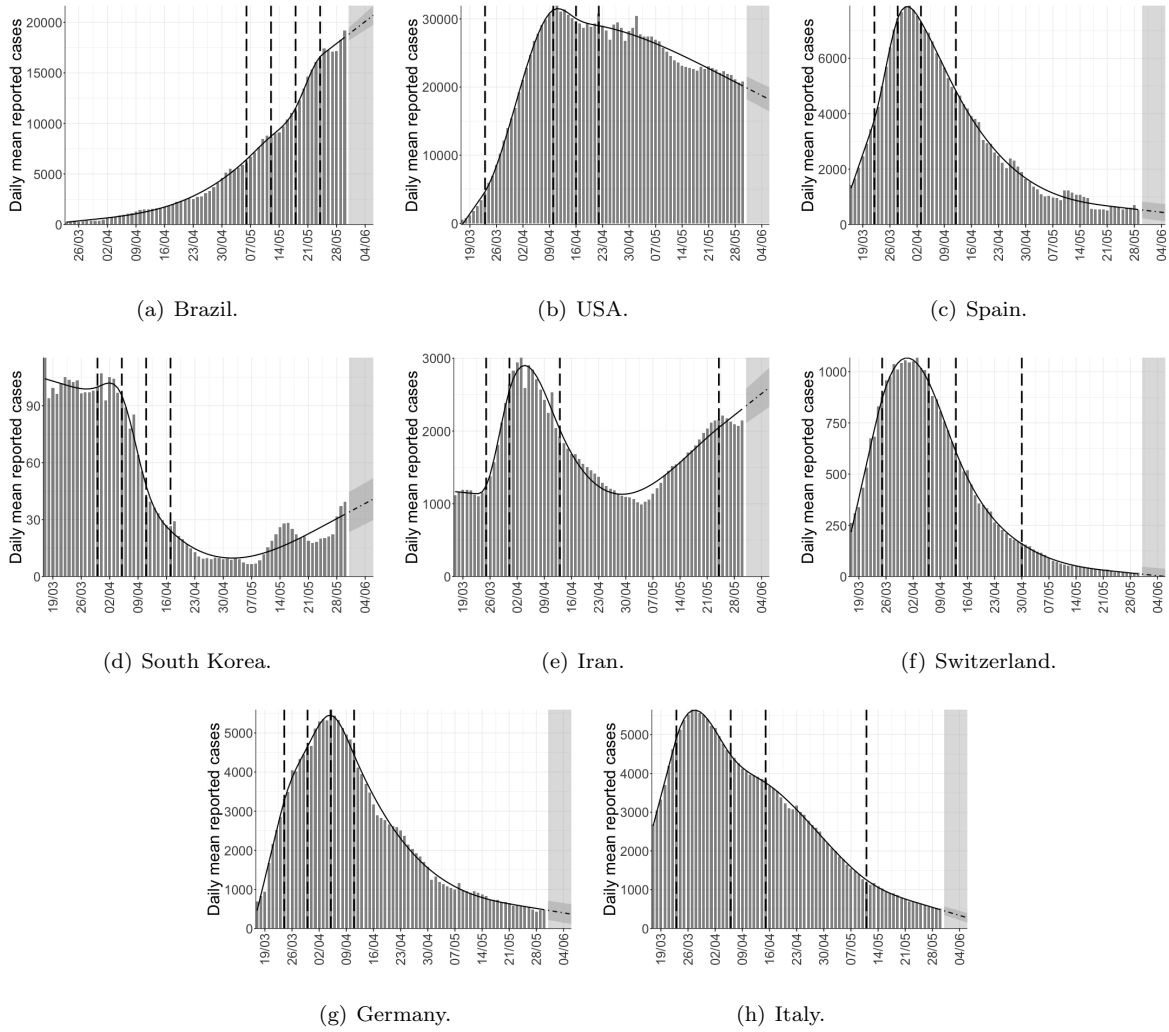
(e) Iran.

(f) Switzerland.

(g) Germany.

(h) Italy.

Figure 4.1: Natural cubic splines curve fitting of mean daily new registered cases of Covid-19 in linear scale for eight countries considering data until May 30th. In each graph, the gray bars indicate mean daily recorded new cases, the solid black line is the fitted curve, dotted black line is the prediction for the following 7 days, the shaded area indicates the interval of estimates, vertical dashed lines represent the knots locations.

(a) Brazil.

(b) USA.

(c) Spain.

(d) South Korea.
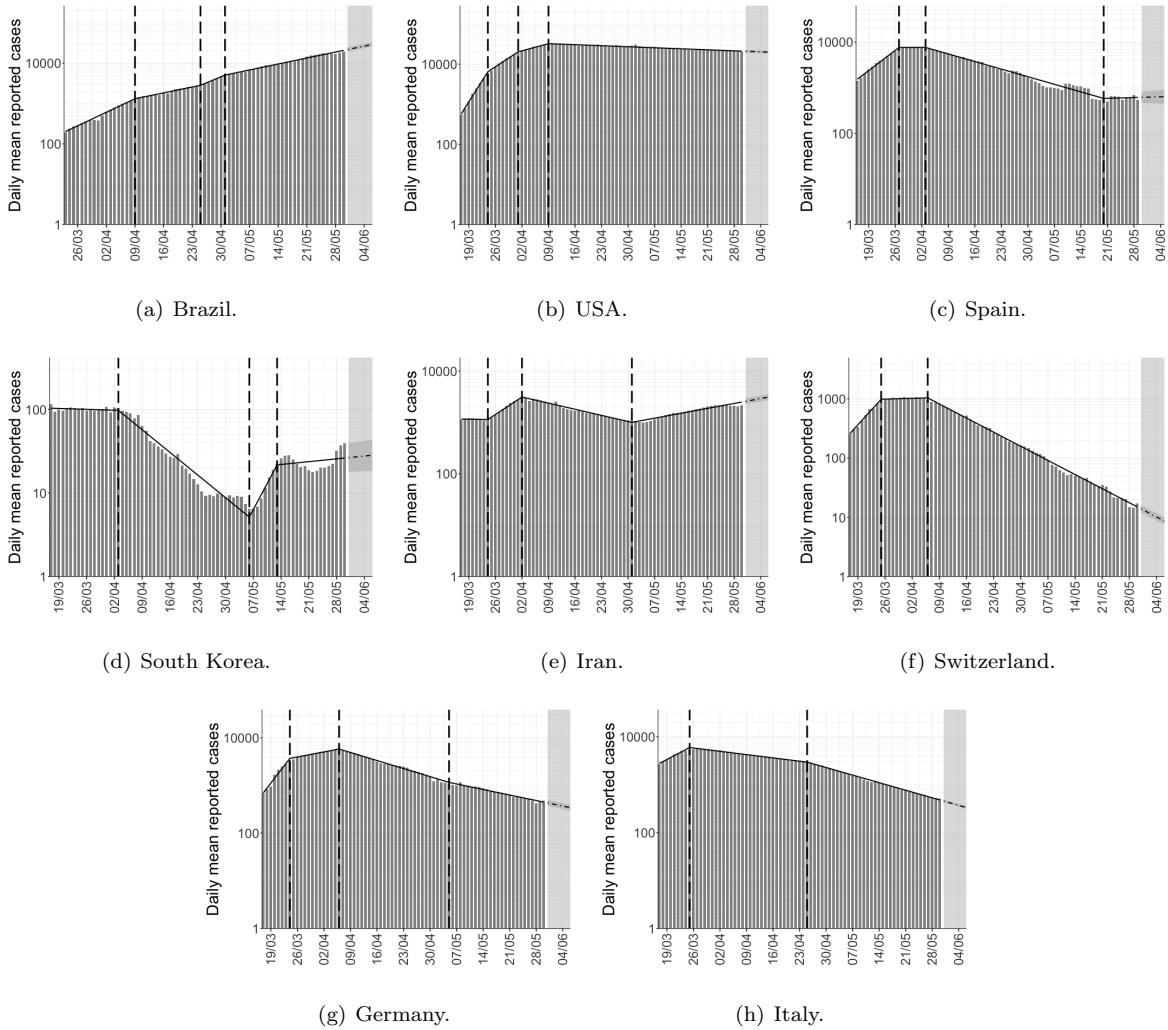
(e) Iran.

(f) Switzerland.

(g) Germany.

(h) Italy.

Figure 4.2: Linear splines curve fitting of mean daily new registered cases of Covid-19 in logarithmic scale for eight countries considering data until May 30th. In each graph, the gray bars indicate mean daily recorded new cases, the solid black line is the fitted curve, dotted black line is the prediction for the following 7 days, the shaded area indicates the interval of estimates, vertical dashed lines represent the knots locations.

5 Discussion

In this paper we introduced a method to estimate the number and position of the knots of a spline regression function. The method is based on the minimization of the sum of squared residuals plus a penalty that depends on the number of knots. We evaluate the performance of the criterion on simulated data, and we showed that our proposed method has a great performance in the simulation studies, even with low SNR values. We also applied the method to perform a descriptive data analysis on covid-19 daily reported cases in several countries. In this analysis we show that the penalized least squares estimation guaranteed quite smooth curve fittings that estimated accurately underlying smooth functions automatically, i.e, without

requiring to specify the number of internal knots and their locations, which is necessary in most spline based curve fitting methods available in the literature.

The penalizing constant λ used in our data analyses was set to a fixed value. However as in many other similar approaches, it could be chosen by a cross validation procedure. From a theoretical point of view, one open question that remains is what is the rate for the penalizing constant λ in order to obtain consistency of the estimator $\hat{\theta}$. Some results on this direction for related penalized models is presented in Castro et al. (2018), where a consistency result was proved for a penalizing constant of order $n^{-1/2}$. Whether this also holds on this setting is an issue that we would like to address in future work.

References

Biller, C. (2000). Adaptive bayesian regression splines in semiparametric generalized linear models. *Journal of Computational and Graphical Statistics*, 9(1):122–140.

Castro, B. M., Lemes, R. B., Cesar, J., Hünemeier, T., and Leonardi, F. (2018). A model selection approach for multiple sequence segmentation and dimensionality reduction. *Journal of Multivariate Analysis*, 167:319 – 330.

De Boor, C. (2001). Calculation of the smoothing spline with weighted roughness measure. *Mathematical Models and Methods in Applied Sciences*, 11(01):33–41.

De Boor, C., De Boor, C., Mathématicien, E.-U., De Boor, C., and De Boor, C. (1978). *A practical guide to splines*, volume 27. springer-verlag New York.

Denison, D., Mallick, B., and Smith, A. (1998). Automatic bayesian curve fitting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(2):333–350.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*. Springer Series in Statistics New York.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The annals of statistics*, pages 1–67.

Friedman, J. H. and Silverman, B. W. (1989). Flexible parsimonious smoothing and additive modeling. *Technometrics*, 31(1):3–21.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. Springer New York.

Osborne, M. R., Presnell, B., and Turlach, B. A. (1998). Knot selection for regression splines via the lasso. *Computing Science and Statistics*, pages 44–49.

Ramsay, J. O. (2004). Functional data analysis. *Encyclopedia of Statistical Sciences*, 4.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of computational and graphical statistics*, 11(4):735–757.

Stone, C. J., Hansen, M. H., Kooperberg, C., Truong, Y. K., et al. (1997). Polynomial splines and their tensor products in extended linear modeling: 1994 wald memorial lecture. *The Annals of Statistics*, 25(4):1371–1470.

Ülker, E. and Arslan, A. (2009). Automatic knot adjustment using an artificial immune system for b-spline curve approximation. *Information Sciences*, 179(10):1483–1494.

Wahba, G. (1990). *Spline models for observational data*, volume 59. Siam.