Statistica Sinica Preprint No: SS-2020-0167	
Title	Dynamic Penalized Splines for Streaming Data
Manuscript ID	SS-2020-0167
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202020.0167
Complete List of Authors	Dingchuan Xue and Fang Yao
Corresponding Author	Fang Yao
E-mail	fyao@math.pku.edu.cn
Notice: Accepted version subject to English editing.	

DYNAMIC PENALIZED SPLINES FOR STREAMING DATA

Dingchuan Xue and Fang Yao

School of Mathematical Sciences, Center for Statistical Science, Peking University

Abstract: We propose a dynamic version of the penalized spline regression designed for streaming data and allows inserting new knots dynamically based on sequential updates of summary statistics. A new theory using direct functional methods rather than traditional matrix analysis is developed to attain the optimal convergence rate in L^2 sense for the dynamic estimation (also applicable for standard penalized splines) under weaker conditions than those in existing work for standard penalized splines.

Key words and phrases: Nonparametric regression, convergence rate, streaming data.

1. Introduction

Penalized spline regression is a computationally efficient method for reconstructing smooth functions from noisy data, which usually starts with a sequence of knots prior to the knowledge of data, then find the spline with given knots that minimize the total squared error plus a penalty on its qth derivate. Specifically, suppose data $\{(x_i, y_i)\}_{i=1,\ldots,n}$ are sampled from a nonparametric model

$$y_i = f_0(x_i) + \varepsilon_i$$

for some unknown function $f_0 : [0,1] \to \mathbb{R}$ contaminated with an independent error ε_i. The penalized spline estimate of f_0 is given by

$$\hat{f}_n = \arg \min_{f \in \mathbb{S}_{\kappa_n, p+1}} \sum_{i=1}^{n} \{y_i - f(x_i)\}^2 + \lambda_n \int_0^1 f^{(q)^2}(x) dx, \qquad (1.1)$$

where $p \geq q$ are positive integers, $\kappa_n = \{0 = \kappa_{n,1} \leq \cdots \leq \kappa_{n,k_n} = 1\} \subseteq [0,1]$ is the set of chosen knots,

$$\mathbb{S}_{\kappa_n, p+1} = \{f \in C^{p-1}([0,1]) : f|_{[\kappa_{n,i}, \kappa_{n,i+1}]} \in \mathbb{P}_p,\ i = 1, \ldots, k_n - 1\} \qquad (1.2)$$

is the space of splines of order p, $\mathbb{P}_p$ is the set of polynomial functions of degree not exceeding p, and λ_n is a positive tuning parameter depending on n. By taking a proper basis of $\mathbb{S}_{\kappa_n,p+1}$, the calculation is reduced to performing a ridge-type regression. This formulation was originally proposed in O'Sullivan (1986) with $q = 2$ and $p = 3$, see Claeskens et al. (2009) for an explicit formulation. Generalized cross-validation proposed by Golub et al. (1979) and Wahba (1990) is often used to choose λ_n. Particularly, if $\lambda_n = 0$, the method is called regression spline. If $\kappa_n = \{x_1 \ldots, x_n\}$ and $p = 2q - 1$, it is called the smoothing spline (Craven and Wahba, 1978). It was offered in de Boor (1978) and Eubank (1999) a general guidance for fitting smoothing splines; see the formulation for the case $q = p$ in Ruppert

(2002), Hall and Opsomer (2005) and Yao and Lee (2008), among others. Our main contribution in this article is to propose a dynamic version of the penalized spline estimtion with theoretical guarantee and specifically designed algorithm for streaming data that allows for an adaptive choice of knots sequence.

It is worth mentioning that, to reach a consistent estimation that approximates a function in an infinite dimensional space, we need to have the number of summary statistics grow as the samples streaming in, which differs from the usual online algorithms. For example, Schifano et al. (2016) proposed online updating techniques for parametric regression problems with constant memory size, and Yang et al. (2010) focused on the online learning of group lasso by updating from previous estimation. By comparison, our approach tackles a nonparametric problem with a sequential updating method, where the memory consumption grows much slower than the sample size.

Owing to its technical challenge, there is no existing work on penalized spline approach orientied towards streaming data. To fill in this gap, we propose a dynamic version of penalized spline estimation, making a sensible modification on the target function by adding a projection to the function space of f on the goodness-of-fit term in the right side of (1.1).

Our algorithm requires only a single iteration of data and allows for an adaptive insertion of knots at a cost of a slight precision loss. We show that under certain conditions, the integrated squared error (i.e., L^2-error) of the dynamic estimation converges at the same rate as the standard penalized spline estimation, $O_p\left\{n^{-2q/(2q+1)}\right\}$, which has not been established for dynamic penalized spline method. This result is derived from a novel technique that lifts the spline space to an infinite dimensional one, which can be seamlessly adopted to the proposed dynamic estimation. By the definition in Stone (1982) or Stone (1980), this rate is asymptotically optimal if $p = q$ and $f_0 \in C^q([0,1])$. It is relevant to mention that Speckman (1985) showed this to be the optimal rate of average mean squared error in an empirical sense. It was pointed out in Golubev and Nussbaum (1990) that this is the minimax rate for f_0 in Sobolev balls, and Huang (2003) obtained similar results for regression splines. If $f_0 \in C^{p+1}([0,1])$ and $p \leq 2q-1$, with a nearly equi-spaced knots condition on κ_n, it is also the convergence rate of the average/empirical mean squared error for a "large" number of knots of the standard penalized spline method as shown in Claeskens et al. (2009). This indicates that the size of κ_n makes little contribution to the result once it is sufficiently large, i.e., exceeding a lower bound depending on f_0 and n. Xiao (2019) extended this result to $C^l([0,1])$ for $q \leq l \leq p$ to obtain L^2

and L^∞ rates, while Schwarz and Krivobokova (2016) established an equivalent kernel theory for penalized splines. It is worth noting that we require weaker conditions to attain the optional rate for the proposed dynamic estimation than those by existing work for standard penalized splines (or the "the large number of knots scenario") , e.g. Claeskens et al. (2009); Xiao (2019), while their work also included theories when the number of knots κ_n and the penalty strength λ_n are small, where the estimation behaves like regression spline.

Nevertheless, in practice it is still meaningful to control the size and location of κ_n for computational efficiency. Various methods were proposed to choose κ_n according to the knowledge of data. For instance, it was suggested in Spiriti et al. (2013) a blind search with golden section adjustment or genetic algorithm for knot selection. Lindstrom (1999) proposed free-knot regression splines with penalty on knots. This type of methods usually involve iterative computations over full data and are not applicable when data come in a streaming manner. Thus a proper choice of κ_n with dynamic updates becomes relevant. It is natural to expect the size of κ_n to grow slowly with n to improve estimation. Intuitively we may insert new knots into existing κ_n as the sample size n grows, behaving like we have a new regressor in ridge-type regression. Hence we propose to modify the target function by adding a projection operator, which sequentially elevates the model dimension.

The rest of the article is organized as follows. We present the proposed dynamic penalized spline estimation with its updating algorithm in Section 2, and offer the corresponding theory that outlines the new technique in Section 3. Numerical studies, including simulated and real data examples, are provided in Section 4, while technical proofs are delineated in the online Supplementary Material.

2. Proposed Methodology and Algorithm

2.1 Dynamic penalized spline estimation

Our goal is to develop a dynamic version of penalized spline estimation that is easy to implement via a sequential updating algorithm with theoretical guarantee. The general setting is that the data are collected in a streaming manner, where the ith incoming data cluster consists of m_i pairs of observations, $\{(x_j, y_j) : j = \sum_{k=1}^{i-1} m_k + 1, \ldots, \sum_{k=1}^{i} m_k\}, i = 1, 2, \ldots$. Since our proposed method and theory remain virtually unchanged for the case of each cluster $m_i = 1$, we present in the sequel this setting for notational convenience. Now suppose that we observe data $\{(x_i, y_i)\}_{i=1,2,\ldots}$ in a streaming fashion (i.e., one by one), following the model

$$y_i = f_0(x_i) + \varepsilon_i$$

for some unknown function $f_0 : [0,1] \to \mathbb{R}$ and an error ε_i. For each n, we denote a knot set $\kappa_n = \{\kappa_{n,1} \le \cdots \le \kappa_{n,k_n}\} \subseteq [0,1]$, depending on $x_1, \ldots, x_{n-1}$, $y_1, \ldots, y_{n-1}$ and κ_{n-1}, such that $\kappa_{n-1} \subseteq \kappa_n$. Let p and q be positive integers satisfying $p \ge q$, and $\mathbb{S}_{\kappa_n,p+1}$ as in (1.2). Let $H^1((0,1))$ be the Sobolev space equipped with the inner product

$$\langle g_1, g_2\rangle_{H^1} = \int_0^1 \{g_1(x)g_2(x) + g_1'(x)g_2'(x)\}\, dx.$$

Let P_n be the orthogonal projection from $H^1(0,1)$ to $\mathbb{S}_{\kappa_n,p+1}$ with respect to this norm. We propose the modification of the standard penalized spline regression in (1.1) as follows,

$$\tilde{f}_n = \arg\min_{f\in\mathbb{S}_{\kappa_n,p+1}} \sum_{i=1}^n \{y_i - P_i f(x_i)\}^2 + \lambda_n \int_0^1 f^{(q)^2}(x)dx. \tag{2.3}$$

Note that the projections $\{P_i\}_{i=1}^n$ serve as a bridge linking the full spline space $\mathbb{S}_{\kappa_n,p+1}$ and the partial one $\mathbb{S}_{\kappa_i,p+1}$, where the squared errors of (x_i, y_i) are evaluated in their own reduced spline spaces in the target function (2.3). With this modification, we show in the sequel that the current penalized spline estimate depends on the previous summary statistics using the same tuning parameter and knots as well as the newly added data, which provides an algorithm for streaming data and is referred to as *dynamic penalized spline estimation*. We shall see that in our asymptotic theory the approximation error introduced by this modification is shown negligible. For theoretical convenience, we let P_i be H^1 projections rather than L^2 type to guarantee boundedness of the derivative of $P_i f$ without loss of generality. Now we describe how the estimation is updated dynamically.

Choose a basis $b_i = (b_{i1}, \ldots, b_{il_i})^{\mathrm{T}}$ of $\mathbb{S}_{\kappa_i,p+1}$ for $i = 1, 2, \ldots$. For $i, j \geq 1$, let C_{ij} be the $l_i \times l_j$ matrix with the value in the uth row and the vth column being $C_{ij,uv} = \langle b_{iu}, b_{jv} \rangle_{H^1}$, and let $Q_{ji} = C_{ji} C_{ii}^{-1}$, then

$$(P_i b_{j1}, \ldots, P_i b_{jl_j})^{\mathrm{T}} = Q_{ji}(b_{i1}, \ldots, b_{il_i})^{\mathrm{T}}, \ i \leq j.$$

For $i \leq j \leq k$, since $P_i = P_i P_j$, we have

$$(P_i b_{k1}, \ldots, P_i b_{kl_k})^{\mathrm{T}} = Q_{kj}(P_i b_{j1}, \ldots, P_i b_{jl_j})^{\mathrm{T}} = Q_{kj} Q_{ji}(b_{i1}, \ldots, b_{il_i})^{\mathrm{T}}.$$

Thus

$$Q_{ki} = Q_{kj} Q_{ji}. \tag{2.4}$$

Suppose $\tilde{f}_n = a_1 b_{n1} + \cdots + a_{l_n} b_{nl_n}$, we have the numerical representation for $\tilde{f}_n$ as

$$(a_1, \ldots, a_{l_n})^{\mathrm{T}} = U_n(\lambda_n) T_n,$$

where $U_n(\lambda_n) = (S_n + \lambda_n D_n)^{-1}$, $S_n = \sum_{i=1}^n Q_{ni} b_i(x_i) b_i(x_i)^{\mathrm{T}} Q_{ni}^{\mathrm{T}}$, $D_n = \int_0^1 b_n^{(q)}(x) b_n^{(q)}(x)^{\mathrm{T}} dx$, and $T_n = \sum_{i=1}^n y_i Q_{ni} b_i(x_i)$. Despite its complicated expression, it is simple to calculate S_{n+1} and T_{n+1} given S_n, T_n, x_{n+1} and y_{n+1}. If $\kappa_{n+1} = \kappa_n$ (no new knots), we may choose that $b_{n+1} = b_n$, then

$$S_{n+1} = S_n + b_{n+1}(x_{n+1})b_{n+1}(x_{n+1})^{\mathrm{T}},\ T_{n+1} = T_n + y_{n+1}b_{n+1}(x_{n+1}).$$

If a new knot is inserted, i.e., $\kappa_{n+1} \supsetneq \kappa_n$, by (2.4) we have

$$\begin{aligned} S_{n+1} &= Q_{n+1,n}S_nQ_{n+1,n}^{\mathrm{T}} + b_{n+1}(x_{n+1})b_{n+1}(x_{n+1})^{\mathrm{T}}, \\ T_{n+1} &= Q_{n+1,n}T_n + y_{n+1}b_{n+1}(x_{n+1}). \end{aligned}$$

With these equations, we are able to update S_n and T_n in an sequential manner. When $\kappa_{n+1} = \kappa_n$ and $\lambda_{n+1} = \lambda_n$, $U_n(\lambda_n)$ can be updated using the Sherman-Morrison formula,

$$U_{n+1}(\lambda_n) = U_n(\lambda_n) - \frac{U_n(\lambda_n)b_{n+1}(x_{n+1})b_{n+1}(x_{n+1})^{\mathrm{T}}U_n(\lambda_n)}{1 + b_{n+1}(x_{n+1})^{\mathrm{T}}U_n(\lambda_n)b_{n+1}(x_{n+1})}.$$

Note that both κ_n and λ_n grows much slower than n, thus in most cases we may update λ_n only when κ_n is changed, which greatly reduces the calculation of matrix inversions.

For the computational complexity, when not inserting new knot or updating λ_n, our update procedure involves only a few matrix-vector multiplication of scale $|\kappa_n|$ that is $O(|\kappa_n|^2)$. The insertion of knots or update of λ_n involves complexity $O(|\kappa_n|^3)$, which occurs on average $O(|\kappa_n|/n)$ of times. Thus the overall computational complexity of the proposed update procedure is $O(|\kappa_n|^2 m + |\kappa_n|^4 m/n)$ for a block of m data points, which is generally much smaller than the complexity $O(|\kappa_n|^2 n)$ of the standard method, where n is the sample size.

2.2 Implementation and dynamic knots insertion

When the tuning parameter λ_n is updated (often together with updating κ_n), it can be tuned by minimizing the generalized cross-validation score. Suppose $(\tilde{f}_n(y_1), \ldots, \tilde{f}_n(y_n))^{\mathrm{T}} = A_n(\lambda_n)(y_1, \ldots, y_n)^{\mathrm{T}}$, the generalized cross-validation score as in Golub et al. (1979) is

$$V(\lambda_n) = \frac{n\left\|\{I - A_n(\lambda_n)\}(y_1, \ldots, y_n)^{\mathrm{T}}\right\|^2}{Tr\{I - A_n(\lambda_n)\}^2}.$$

This can be rewritten as

$$\frac{n\left\{R_n + T_n^{\mathrm{T}} U_n(\lambda_n) S_n U_n(\lambda_n) T_n - 2 T_n^{\mathrm{T}} U_n(\lambda_n) T_n\right\}}{\left[n - Tr\{S_n U_n(\lambda_n)\}\right]^2}, \tag{2.5}$$

where $R_n = \sum_{i=1}^n y_i^2$.

The set of knots κ_{n+1} can be updated with various algorithms. As an example, we use the following method in our implementation, while other methods are also viable as long as they can be updated dynamically for streaming data. The theory in Theorem 2 suggests that we may let $\kappa_{n+1} = \kappa_n$ for most n, which is in accordance to the intuition that knots grow slowly relative to sample size. We introduce a parameter ν that reflects the spanning of κ_n, i.e. $E\Delta_n = O(n^{-\nu})$ with $\Delta_n = \max_j |\kappa_{n,j} - \kappa_{n,j+1}|$. Our theory implies that, given $\nu > (2q-1)/\{(2q+1)(2q-3)\}$ and $\alpha > 0$, we may add new knots when $n > \alpha|\kappa_{n-1}|^{1/\nu}$. If we are to insert a new knot x into κ_n such that $\kappa_{n+1} = \kappa_n \cup \{x\}$, we put x in a similar way to that in Yuan and Zhou (2012). According to Proposition 6, Section 1.5.3.2 in Kunoth et al. (2017),

$$\inf_{s \in \mathbb{S}_{\kappa_n, p+1}} \|f_0 - s\|_{L^2([\kappa_{n,i}, \kappa_{n,i+1}])} \leq K\left(\kappa_{n,i+p+1} - \kappa_{n,i-p}\right)^q \left\| f_0^{(q)} \right\|_{L^2([\kappa_{n,i-p}, \kappa_{n,i+p+1}])}$$

for some constant K. We suggest to insert the new point where this bound is large, with f_0 replaced by $\tilde{f}_n$. Let

$$j = \arg\max_j \left(\kappa_{n,j+p+1} - \kappa_{n,j-p}\right)^q \left\| \tilde{f}_n^{(q)} \right\|_{L^2([\kappa_{n,j-p}, \kappa_{n,j+p+1}])}, \tag{2.6}$$

then a new knot is placed at $(\kappa_{n,i} + \kappa_{n,i+1})/2$, where

$$i = \arg\max_{j-p \leq i \leq j+p} (\kappa_{i+1} - \kappa_i). \tag{2.7}$$

This is a light weighted algorithm compared to the matrix algebraic computations. Such way of selecting new knots tends to place more knots where the curve changes sharply. The limiting behavior of the algorithm would have the density of knots roughly proportional to $|f_0^{(q)}(x)|^{1/q}$.

We summarize the proposed dynamic penalized spline estimation into the algorithm as follows. Given an initial knot sequence κ_0, the spline order p and the penalty order q, the values of ν and α for knot insertion, let $\{b_{0,1}, \ldots, b_{0,l_0}\}$ be a basis of $\mathbb{S}_{\kappa_0,p+1}$. Let S_0, T_0 and R_0 be zeroes in $\mathbb{R}^{l_0 \times l_0}$, $\mathbb{R}^{l_0}$ and $\mathbb{R}$, and $R_n = \sum_{i=1}^n y_i^2$.

In practice, the parameter ν can be chosen to be slightly larger than its theoretical bound $(2q-1)/\{(2q+1)(2q-3)\}$ given in Theorem 2, and α can be tuned with the first batch of samples to achieve a balance between the number of knots and the generalized cross-validation scores, as shown in our numerical studies. Moreover, after one chooses α this way, the resulting estimates are fairly stable when varying the value of ν under the constraint $\alpha|\kappa_{n-1}|^{1/\nu} < n$. This provides a practical guidance of choosing ν and α given the penalty order q. We conclude this section by noting that the proposed method and algorithm, as well as the theory in next section, can be straightforwardly extended to the case of multivariate covariates with slight modification.

3. Theoretical Results

Before stating the main result, we give a corresponding result on L^2 convergence of standard penalized spline that has not been attained in literature. The proof is deferred to the Supplementary Material, in which the techniques are useful in analyzing the dynamic penalized splines. A standard

for $n = 1, 2, \ldots$ **do**

if $n > \max\{\alpha|\kappa_{n-1}|^{1/\nu}, p\}$ **then**

Let κ_* be the new knot as defined in (2.6) and (2.7) and

$\kappa_n = \kappa_{n-1} \cup \{\kappa_*\}$;

Choose a basis $b_n = (b_{n,1}, \ldots, b_{n,l_n})^{\mathrm{T}}$ for $\mathbb{S}_{\kappa_n,p+1}$;

Let $C_{n-1,n-1}$ be the matrix that

$C_{n-1,n-1,uv} = (b_{n-1,u}, b_{n-1,v})_{H_1}$;

Let $C_{n,n-1}$ be the matrix that $C_{n,n-1,uv} = (b_{n,u}, b_{n-1,v})_{H_1}$;

Let $Q_{n,n-1} = C_{n,n-1}C^{-1}_{n-1,n-1}$;

Let $S_n = Q_{n,n-1}S_{n-1}Q^{\mathrm{T}}_{n,n-1} + b_n(x_n)b_n(x_n)^{\mathrm{T}}$,

$T_n = Q_{n,n-1}T_{n-1} + y_n b_n(x_n)$ and $R_n = R_{n-1} + y_n^2$;

else

Let $\kappa_n = \kappa_{n-1}$ and $b_n = b_{n-1}$;

Let $S_n = S_{n-1} + b_n(x_n)b_n(x_n)^{\mathrm{T}}$, $T_n = T_{n-1} + y_n b_n(x_n)$ and

$R_n = R_{n-1} + y_n^2$;

end

Let $D_n = \int_0^1 b_n^{(q)}(x)b_n^{(q)}(x)^{\mathrm{T}}dx$ and λ_n be the minimizer of (2.5);

Let $\tilde{f}_n(x) = b_n(x)^{\mathrm{T}}(S_n + \lambda_n D_n)^{-1}T_n$;

end

condition below is imposed for the penalized spline estimation defined in (1.1).

Assumption 1. $f_0 \in C^l([0,1])$ for some $l \geq q$ or $f_0 \in H^l([0,1])$ for some $l \geq q+1$, $p \geq q \geq 2$, where $H^l([0,1])$ is the Sobolev space slightly larger than C^l.

Recall that $\Delta_i = \max_{1\leq j\leq k_i} |\kappa_{i,j+1} - \kappa_{ij}|$, let $F_i(x) = \sum_{j=1}^{i} \mathbf{1}_{x\geq x_j}/i$, $E_j(x) = \sum_{j=1}^{i} \mathbf{1}_{x\geq x_j}\varepsilon_j$ and $M_j = \max_{0\leq x\leq 1} E_j(x)$, where $\mathbf{1}_{x\geq x_j}$ is 1 when $x \geq x_j$ and 0 otherwise. We suppose F_n converges to some differentiable function F. To be precise,

Assumption 2. F is a continuously differentiable probability distribution function on $[0,1]$, such that $0 < \min_x F'(x) \leq \max_x F'(x) < \infty$.

Assumption 3. $\|F_n - F\|_\infty = O_p\left(n^{-1/2}\right)$ and $M_n = O_p\left(n^{1/2}\right)$.

When $x_1, x_2, \ldots$ are independently and identically distributed from the distribution F, it is well-known that $\|F_n - F\|_\infty = O_p\left(n^{-1/2}\right)$. Furthermore, when $\varepsilon_1, \varepsilon_2, \ldots$ are zero-mean and independent (also independent of $x_1, x_2, \ldots$) with second moment uniformly bounded by M, from Doob's martingale inequality, one has $P(M_n \geq \alpha) \leq (nM)^{1/2}/\alpha$ for all $\alpha > 0$, which implies $M_n = O_p\left(n^{1/2}\right)$. For non-random $x_1, x_2, \ldots$, this assumption simply correspond to its non-random version $\|F_n - F\|_\infty = O\left(n^{-1/2}\right)$

and $M_n = O\left(n^{1/2}\right)$. When working with large number of knots, that is, the "smoothing spline" scenario in Claeskens et al. (2009), unlike existing theories for penalized spline, we impose neither an explicit assumption on the distributions of x_i or y_i, nor a lower bound on the distance between adjacent knots in κ_n (e.g., Claeskens et al., 2009).

Theorem 1. *Given Assumptions 1 and 2, there exist constants C_1, C_2 depending on l, p, q, f_0, F, when the following holds,*

$$\|F_n - F\|_\infty \lambda_n^{-\frac{1}{2q}} n^{\frac{1}{2q}} \le C_1, \quad \lambda_n \le C_1 n, \tag{3.8}$$

we have

$$\left\|f_0 - \hat{f}_n\right\|_2^2 \le C_2 \Delta_n^{2min\{l,p+1\}} + C_2\lambda_n/n + C_2 M_n^2 \lambda_n^{-\frac{1}{2q}} n^{-\frac{4q-1}{2q}}, \tag{3.9}$$

where $\hat{f}_n$ is the standard penalized spline estimation defined in (1.1).

If we additionally impose Assumption 3, then for $D_1 n^{1/(2q+1)} \le \lambda_n \le D_2 n^{1/(2q+1)}$, $D_1, D_2 \in (0, \infty)$ and $\Delta_n = O_p\left\{(\lambda_n/n)^{1/(2min\{l,p+1\})}\right\}$, we have

$$\left\|f_0 - \hat{f}_n\right\|_2^2 = O_p\left(n^{-\frac{2q}{2q+1}}\right).$$

The inequality (3.9) reveals the relation between λ_n/n and $\Delta_n^{2\min\{l,p+1\}}$. For instance, if $(\lambda_n/n)^{-1/(2\min\{l,p+1\})} \ge C|\kappa_n|$ for some C, the first term $\Delta_n^{2\min\{l,p+1\}}$ shall dominate, which is usually not desired.

Compared to the conditions assumed in Claeskens et al. (2009), this L^2 convergence rate does not require a lower bound of $\min_i |\kappa_{n,i+1} - \kappa_{n,i}|$. In the second part of the theorem, Assumption 3 and $D_1 n^{1/(2q+1)} \le \lambda_n \le D_2 n^{1/(2q+1)}$ together implies (3.8) by noting

$$\|F_n - F\|_\infty \lambda_n^{-\frac{1}{2q}} n^{\frac{1}{2q}} = O_p(n^{\frac{1-2q}{4q+2}}), \qquad \lambda_n = o(n).$$

Stone (1982) has shown that under certain conditions, if (x_i, y_i) are simple random samples with $Ey_i = f_0(x_i)$ and $l = q$, the rate $O_p\left\{n^{-2q/(2q+1)}\right\}$ is optimal for integrated squared error. With stronger assumptions, Claeskens et al. (2009) showed the convergence rate of average mean squared error (in an empirical sense), $\sum_{i=1}^n \{f_0(x_i) - \hat{f}_n(x_i)\}^2/n = O_p\left\{n^{-2q/(2q+1)}\right\}$, for a large number of knots, and $O_p\left\{n^{-(2p+2)/(2p+3)}\right\}$ for a small number of knots. Such results were attained under a stronger condition that, roughly speaking, knots in κ_n are not far from equi-spaced.

Next we present the result for the proposed dynamic penalized spline estimation, and requires additional assumptions as follows.

Assumption 4. $\sup_{i=1,2,\ldots} E\varepsilon_i^2 < \infty$, $E\varepsilon_i = 0$ for $i = 1, 2, \ldots$. Either $\{\varepsilon_i\}_{i=1,2,\ldots}$ are pairwise uncorrelated and independent of $\{\kappa_i\}_{i=1,2,\ldots}$ and $\{x_i\}_{i=1,2,\ldots}$, or $\{\varepsilon_i\}_{i=1,2,\ldots}$ are pairwise independent and ε_j is independent of κ_i and x_i for $i \le j$.

Assumption 5. $D_1 n^{1/(2q+1)} \leq \lambda_n \leq D_2 n^{1/(2q+1)}$ for some $D_1, D_2 \in (0, \infty)$, $E\mathit{\Delta}_n = O\left(n^{-\nu}\right)$, $\|F_n - F\|_\infty^2 |\kappa_{2n+1}| = o_p\left(n^\xi\right)$ and $\sum_{j\leq n:\kappa_{j+1}\neq\kappa_j} \|F_j - F\|_\infty^2 = o_p\left(n^\xi\right)$ for some $\nu > (2q-1)/\{(2q+1)(2q-3)\}$ and $\xi = (2q-2)\nu + 2q/(2q+1)$.

Assumption 4 is a rather mild condition and is apparently satisfied by most situations where x_i's and κ_i's are commonly assumed independent ε_i's. Assumption 5 imposes conditions on the distribution of x_i's and the growth of κ_n, where the spanning $\mathit{\Delta}_n$ on average is assumed at a polynomial order of n. The conditions $\|F_n - F\|_\infty^2 |\kappa_{2n+1}| = o_p\left(n^\xi\right)$ and $\sum_{j\leq n:\kappa_{j+1}\neq\kappa_j} \|F_j - F\|_\infty^2 = o_p\left(n^\xi\right)$ are actually implied by a stronger one, $D_3 n^\nu \leq |\kappa_n| \leq D_4 n^\nu$, that was adopted in most existing work on standard spline estimation (e.g. Claeskens et al., 2009; Wang et al., 2011; Schwarz and Krivobokova, 2016; Xiao, 2019). Note that the condition $\|F_n - F\|_\infty^2 |\kappa_{2n+1}| = o_P(n^\xi)$ is different from $\|F_n - F\|_\infty^2 |\kappa_n| = o_P(n^\xi)$. Roughly speaking, this assumption requires that the distribution pattern of later samples to not differ dramatically from that of the early ones.

Theorem 2. *Suppose that Assumptions 1–5 hold, then we have*

$$\left\|f_0 - \tilde{f}_n\right\|_2^2 = O_p\left(n^{-\frac{2q}{2q+1}}\right),$$

where $\tilde{f}_n$ is the dynamic penalized spline as defined in (2.3).

Note that the results holding in probability is a consequence of the random design points $\{x_i\}$. Our assumptions on F_n are in the form of O_P or o_P, which is the usual case for i.i.d. design points. Had those assumption be replaced with nonrandom uniform bounds, the reader may follow our proof and arrive at similar results of $E\left\|f_0 - \tilde{f}_n\right\|^2$.

Distinct from Hall and Opsomer (2005), Claeskens et al. (2009) and Xiao (2019) which built their arguments on the analysis of matrices, our proof deals directly with function spaces, which provides a new and general technique that is sketched below.

Our theory has an origin from Munteanu (1973), which is adopted for penalized splines. Let Z be the Hilbert space $L^2 \times \mathbb{R}^n$, with the inner product defined by

$$\langle (g_1, z_{11}, \ldots, z_{1n}), (g_2, z_{21}, \ldots, z_{2n}) \rangle_Z = \lambda_n \int_0^1 g_1(x) g_2(x) dx + \sum_{i=1}^n z_{1i} z_{2i}.$$

Let $L : H^q \to Z$ be the bounded linear map given by

$$Lg = \left(g^{(q)}, P_1 g(x_1), \ldots, P_n g(x_n)\right).$$

We show that

$$\sup_g \|g\|_2^2 / \|Lg\|_Z^2 = O_p\left(n^{-1}\right) \tag{3.10}$$

and

$$\left\|Lf_0 - L\tilde{f}_n\right\|_Z^2 = O_p\left\{n^{1/(2q+1)}\right\}. \tag{3.11}$$

The first part (3.10) is done by showing that

$$\sup_g \frac{n\|g\|_2^2 + \lambda_n \left\|g^{(q)}\right\|_2^2 - \|Lg\|_Z^2}{n\|g\|_2^2 + \lambda_n \|g^{(q)}\|_2^2} = o_p(1).$$

For (3.11), let $h = (0, y_1, \ldots, y_n) \in Z$, and let $Q_1 : Z \to LH^q$ and $Q_2 : Z \to L\mathbb{S}_{\kappa_n, p+1}$ be orthogonal projection, then $L\tilde{f}_n = Q_2 h$ and $Q_2 = Q_2 Q_1$. We have that

$$\begin{aligned} \left\|Lf_0 - L\tilde{f}_n\right\|^2 &= \|Lf_0 - Q_2 Lf_0\|^2 + \left\|Q_2 Lf_0 - L\hat{f}_n\right\|^2 \\ &\leq \|Lf_0 - Q_2 Lf_0\|^2 + \|Q_1 Lf_0 - Q_1 h\|^2 . \end{aligned}$$

From the theory of splines in Schumaker (2007), there exists $s \in \mathbb{S}_{\kappa_n, p+1}$ and $C > 0$ such that

$$\left\|f_0^{(r)} - s^{(r)}\right\|_q \leq C\Delta^{l-r} \left\|f_n^{(l)}\right\|_q, \quad 0 \leq r \leq l-1,$$

thus

$$\|Lf_0 - Q_2 Lf_0\|^2 \leq \{1 + o_p(1)\} \left(n\|f_0 - s\|_2^2 + \lambda_n \left\|f_0^{(q)} - s^{(q)}\right\|_2^2 \right) = O_p\left\{n^{1/(2q+1)}\right\}.$$

We may also show $\|Q_1 Lf_0 - Q_1 h\|^2 = O_p\left\{n^{1/(2q+1)}\right\}$ from the fact that

$$\|Q_1 Lf_0 - Q_1 h\| = \sup_{g \in H^q} \frac{\langle Lg, Lf_0 - h\rangle_Z}{\|Lg\|}.$$

A detailed proof is given in the online Supplementary Material, while the proof for the standard penalized spline estimation is to substitute the definition of L with $Lg = \left(g^{(q)}, g(x_1), \ldots, g(x_n)\right)$.

4. Numerical Study

4.1 Simulated examples

We generate independent $x_1, x_2, \ldots$ and $\varepsilon_1, \varepsilon_2, \ldots$ in simulation studies. For the first example, let x_i be uniformly distributed on $[0, 1]$, ε_i follow the standard normal distribution $N(0, 1)$, and $f_0(x) = 50(x-0.5)\exp\{-100(x-0.5)^2\}$.

We consider fitting this model with two smoothness/penalty settings, $p = 3, q = 2$ or $p = 4, q = 3$. Starting with an initial $\kappa_1 = \{0, 0.2, 0.4, 0.6, 0.8, 1\}$, we take $\nu = 2/3$ for the former setting, and $\nu = 1/3$ for the latter. We evaluate the performance of the dynamic and standard penalized spline estimation with various values of α, and the total sample size is 5×10^4. We calculate the bias, variance and total mean squared error, denoted by $L^2_{bias} = \|f_0 - E\tilde{f}_n\|_2^2$, $L^2_{var} = E\|\tilde{f}_n - E\tilde{f}_n\|_2^2$ and $L^2_{err} = E\|f_0 - \tilde{f}_n\|_2^2$, by averaging over 1000 Monte Carlo runs. The results are shown in the Table 1, which indicates the dynamic penalized estimation performs as well as the standard method, no matter whether one used the common equi-spaced knots or the knots chosen by the dynamic method (the knots size equals to $|\kappa_n|$). This provides empirical support that the potential precision loss caused by modifying the target function (1.1) is numerically negligible. Note that we fixed ν slightly larger than $(2q-1)/\{(2q+1)(2q-3)\}$ in each smooth/penalty setting, the estimation with different values of α appears fairly stable. It is worth mentioning that the dynamic updates need only the previous-step estimates using newly added data.

To see the influence of α and ν, we first fix ν slightly larger than its theoretical lower bound as above, and tune α with the first batch of samples. Figure 1 shows the generalized cross-validation scores versus different values of α for the first 500, 1000 and 1500 samples, respectively. We see that $\alpha = 2$ appears to reasonably balance the knots size and performance for $p = 3$, $q = 2$ and $\nu = 2/3$, as a larger α encourages fewer knots and potentially elevates the estimation error. Analogously, we may choose $\alpha = 0.04$ for the case of $p = 4$, $q = 3$ and $\nu = 1/3$. It is also seen that, the number of samples has little impact on the choice of α when it is adequate. Moreover, with this selected α, the influence on the generalized cross-validation score from the choice of ν is fairly minor, shown in Fig. 2. This provides empirical support on how to choose ν and α in practice, and the performance is rather stable in a wide range of α (and ν).

Our method and theory can be naturally extend to modeling multi-dimensional y_i, and the algorithm for choosing new knots remains unchanged. In the second example, we let y_i be a bivariate response. With $f_0(x) = (g(x)\sin x, g(x)\cos x)^{\mathrm{T}}$, where $g(x) = (2\pi x + 20\pi x^3)/(1 + x^3)$, ε_i

Table 1: Results of our first simulated example with the total sample size 5×10^4. The abbreviation DS stands for the proposed dynamic penalized estimation, PS_1 for the standard penalized spline estimation with λ_n tuned by generalized cross-validation and the knots equi-spaced on $[0, 1]$ with the size equal to $|\kappa_n|$ of the dynamic method, and PS_2 for the standard penalized spline estimation with the knots κ_n from the dynamic method. Shown are the Monte Carlo averages over 1000 runs for $L^2_{bias} = \|f_0 - E\tilde{f}_n\|_2^2$, $L^2_{var} = E\|\tilde{f}_n - E\tilde{f}_n\|_2^2$ and $L^2_{err} = E\|f_0 - \tilde{f}_n\|_2^2$, all multiplied by 10^4 for visualization.

p, q, ν	α	L^2_{bias}			L^2_{var}			L^2_{err}		
		DS	PS_1	PS_2	DS	PS_1	PS_2	DS	PS_1	PS_2
$3, 2, 2/3$	1	2.25	2.26	2.26	18.9	18.9	18.9	21.1	21.2	21.2
	2	2.13	2.16	2.16	18.7	18.6	18.6	20.9	20.8	20.8
	4	2.29	2.36	2.36	18.8	18.5	18.5	21.1	20.9	20.9
$4, 3, 1/3$	.02	1.38	1.39	1.39	17.2	17.2	17.1	18.6	18.6	18.5
	.04	1.29	1.28	1.27	17.1	17.1	17.1	18.4	18.4	18.3
	.08	1.24	1.27	1.23	17.4	17.3	17.3	18.6	18.6	18.5

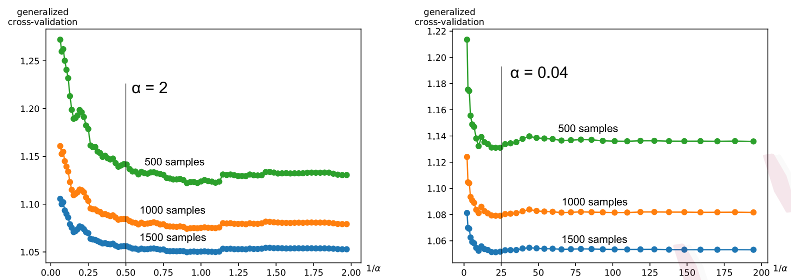

Figure 1: Generalized cross-validation scores of the first batch of samples in one Monte Carlo run with various values of α. For the left panel, $p = 3$, $q = 2$ and $\nu = 2/3$; for the right panel, $p = 4$, $q = 3$ and $\nu = 1/3$.

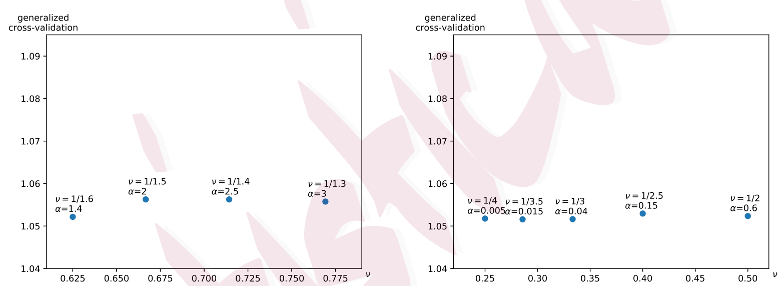

Figure 2: Generalized cross-validation scores of the first 1500 samples in one Monte Carlo run with various values of ν, where the parameter α is tuned as in Fig. 1. For the left panel, $p = 3$ and $q = 2$, where ν is subject to a lower bound constraint at $3/5$. For the right panel, $p = 4$ and $q = 3$, where the lower bound constraint is $5/21$.

follows the bivariate standard normal distribution, and other parameters are the same as in the first example. The penalized spline estimation is performed in two fittings, where the smoothness/penalty parameters (and associated values of ν and α) are given by $p = 3, q = 2, \nu = 2/3, \alpha = 100$ and $p = 4, q = 3, \nu = 1/3, \alpha = 0.4$ respectively, and the total sample size 5×10^4. To appreciate the influence of the knot placement offered by the dynamic estimation, we compare the proposed to the standard method using equispaced knots with the same knots size equal to $|\kappa_n|$. For the first setting, L^2_{err} averaged over 1000 Monte Carlo runs for our and standard methods are 1.563×10^{-3} and 1.530×10^{-3}, respectively, where both bias and variance are similar. For the second setting, we have the L^2_{err} of 1.51×10^{-3} from dynamic estimation ($L^2_{bias} = 2.46 \times 10^{-4}$ and $L^2_{var} = 1.26 \times 10^{-3}$, respectively), and 2.59×10^{-3} from the standard estimation ($L^2_{bias} = 1.48 \times 10^{-3}$ and $L^2_{var} = 1.11 \times 10^{-3}$, respectively). As shown in Fig. 3, for the first setting, the dynamic estimation is close to the standard estimation. For the second, our method seems to put more knots at large values of x with high curvature, which reduced the approximation bias substantially but sufferred slightly larger variance. We also report in Table 2 the average computation time of each single update of our algorithm on our computer with an Intel i5-6500 CPU, which is much faster compared to that of the

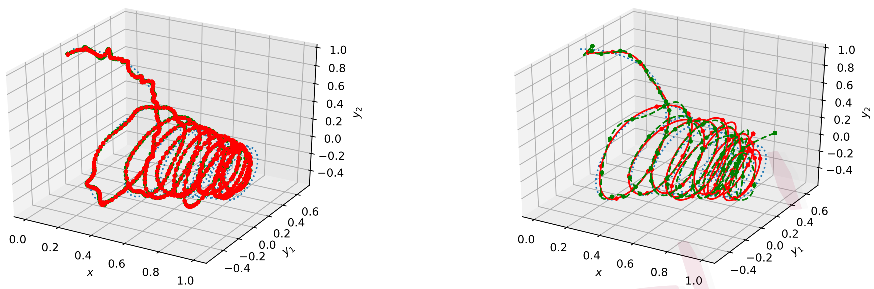

Figure 3: A Monte Carlo run of the second simulated example. The left panel is under the setting $p = 3, q = 2, \nu = 2/3, \alpha = 100$, and the right one is under the setting $p = 4, q = 3, \nu = 1/3, \alpha = 0.4$. The red solid line is the proposed dynamic estimation, the green dash line is the estimation of standard penalized spline estimation with equi-spaced knots of size $|\kappa_n|$, and the blue dotted line is the underlying f_0.

standard penalized spline estimation using a full sample of $n = 1500$ for empirical illustration.

4.2 A real example

We present an application to the regression of power plant output. The dataset comes from Tüfekci (2014), containing 9568 data points collected from a Combined Cycle Power Plant over 6 years 2006–2011, when the power plant was set to work with full load. The features include ambi-

Table 2: Computation time comparison in various settings with sample size $n = 1500$ for illustration. We listed the average time of a single update on our computer with an Intel i5-6500 CPU, and the time of a full computation of the standard penalized spline estimation, both in milliseconds.

p, q, ν	α	Avg. update time(ms)	Std. method(ms)
	1	0.8	24
$3, 2, 2/3$	2	0.5	19
	4	0.3	6
	0.02	0.2	19
$4, 3, 1/3$	0.04	0.2	14
	0.08	0.2	13

ent temperature (AT) measured in whole degrees in Celsius and full load electrical power output (PE) measured in mega watts, shown in Fig 4(a).

We perform a penalized spline regression with the proposed dynamic method and the standard method measuring $E(PE|AT)$, where x_i is the AT of the ith observation, scaled to $[0, 1]$, and y_i is the PE of the ith observation. We perform the regression with two settings, $q = 2$, $p = 3$, $\nu = 2/3$ and $q = 3$, $p = 4$, $\nu = 1/3$. We first obtain estimation with various α on 500 data points, shown in (b) and (d) of Fig 4. From the generalized cross-validation scores we see that $\alpha = 2$ (or 0.125) is an adequate choice for adding knots in the first (or the second) setting. Then we carry out the proposed and the standard methods on the full dataset, denoting the estimates by $\tilde{f}$ and $\hat{f}$ (with the same number of knots as ours but equi-spaced on $[0, 1]$), respectively. We measure the relative L^2 difference between $\tilde{f}$ and $\hat{f}$, $\|\tilde{f} - \hat{f}\|_2/\|\hat{f}\|_2$, which is 1.268×10^{-4} for the first setting and 8.478×10^{-5} for the second. This suggests little difference using the dynamic updates in a streaming manner, compared to the standard estimation using the full data. We also perform a 10-fold cross-validation measuring average mean squared prediction error, which has nearly identical results for dynamic and standard estimation in both settings (not reported for conciseness) . This empirically supports our theory for the dynamic penalized splines, and is also graphically demonstrated in Figure 4 (c) and (e) that the estimates obtained by two methods are visually indistinguishable.

Supplementary Material

The auxiliary lemmas and the proofs to the main theorems are deferred to the online Supplementary Material.

Acknowledgements

Fang Yao is the corresponding author. This research is partially supported by National Natural Science Foundation of China Grants 11931001 and 11871080, the LMAM, and the Key Laboratory of Mathematical Economics and Quantitative Finance (Peking University), Ministry of Education.

References

Claeskens, G., T. Krivobokova, and J. D. Opsomer (2009). Asymptotic properties of penalized spline estimators. *Biometrika 96*(3), 529–544.

Craven, P. and G. Wahba (1978). Smoothing noisy data with spline functions. *Numerische Mathematik 31*(4), 377–403.

de Boor, C. (1978). *A Practical Guide to Splines.* Applied Mathematical Sciences. Springer-Verlag.

Eubank, R. L. (1999). *Nonparametric Regression and Spline Smoothing.* CRC press.

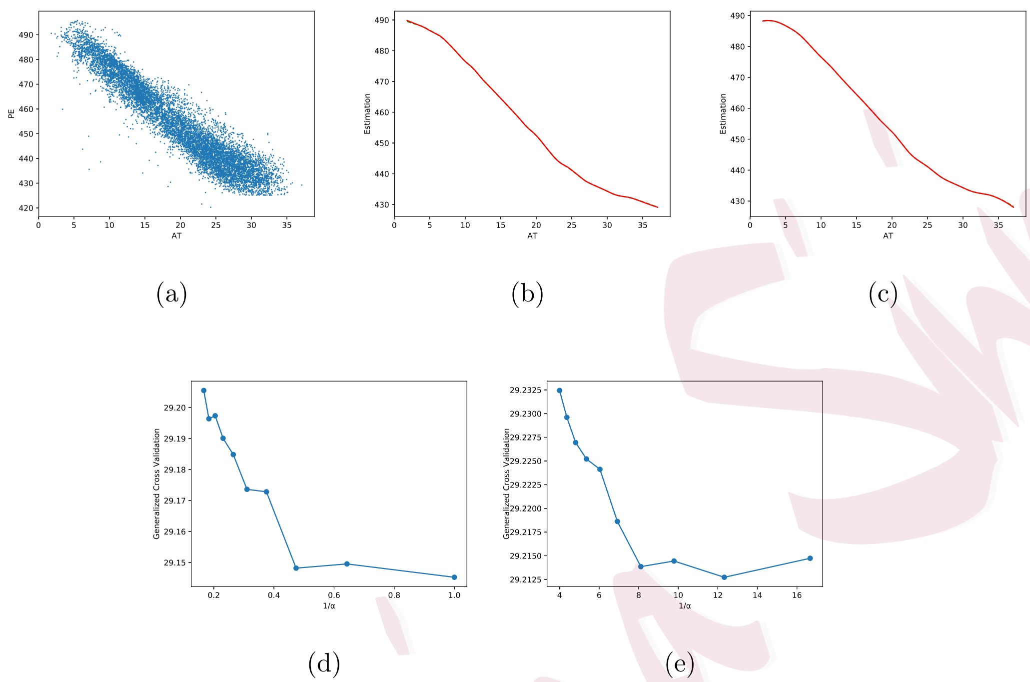

Figure 4: Illustration of the power plant dataset. Panels (b) and (d) are plotted under setting $q = 2$, $p = 3$ and $\nu = 2/3$, while (c) and (e) are plotted under setting $q = 3$, $p = 4$, $\nu = 1/3$. (a): Scatter plot of the dataset. (b) and (c): The red solid line obtained by the proposed method and the green dashed line by the standard estimation are visually indistinguishable. (d) and (e): Generalized cross-validations scores of our method performed on 500 from a total of 9568 sample points with various α, suggesting $\alpha = 2$ and $\alpha = .125$, respectively.

Golub, G. H., M. Heath, and G. Wahba (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics 21*(2), 215–223.

Golubev, G. K. and M. Nussbaum (1990). A risk bound in Sobolev class regression. *The Annals of Statistics 18*(2), 758–778.

Hall, P. and J. D. Opsomer (2005). Theory for penalised spline regression. *Biometrika 92*(1), 105–118.

Huang, J. Z. (2003). Asymptotics for polynomial spline regression under weak conditions. *Statistics & Probability Letters 65*(3), 207–216.

Kunoth, A., T. Lyche, G. Sangalli, and S. Serra-Capizzano (2017). *Splines and PDEs: From Approximation Theory to Numerical Linear Algebra*, Volume 2219 of *Lecture Notes in Mathematics*. Springer.

Lindstrom, M. J. (1999). Penalized estimation of free-knot splines. *Journal of Computational and Graphical Statistics 8*(2), 333–352.

Munteanu, M.-J. (1973). Generalized smoothing spline functions for operators. *SIAM Journal on Numerical Analysis 10*(1), 28–34.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems. *Statistical Science 1*(4), 502–518.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics 11*(4), 735–757.

Schifano, E. D., J. Wu, C. Wang, J. Yan, and M.-H. Chen (2016). Online updating of statistical inference in the big data setting. *Technometrics 58*(3), 393–403.

Schumaker, L. (2007). *Spline Functions: Basic Theory* (3 ed.). Cambridge Mathematical Library. Cambridge University Press.

Schwarz, K. and T. Krivobokova (2016). A unified framework for spline estimators. *Biometrika 103*(1), 121–131.

Speckman, P. (1985). Spline smoothing and optimal rates of convergence in nonparametric regression models. *The Annals of Statistics 13*(3), 970–983.

Spiriti, S., R. Eubank, P. W. Smith, and D. Young (2013). Knot selection for least-squares and penalized splines. *Journal of Statistical Computation and Simulation 83*(6), 1020–1036.

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *The Annals of Statistics 8*(6), 1348–1360.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics 10*(4), 1040–1053.

Tüfekci, P. (2014). Prediction of full load electrical power output of a base load operated combined cycle power plant using machine learning methods. *International Journal of Electrical Power & Energy Systems 60*, 126 – 140.

Wahba, G. (1990). *Spline Models for Observational Data*. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics.

Wang, X., J. Shen, and D. Ruppert (2011). On the asymptotics of penalized spline smoothing. *Electronic Journal of Statistics 5*, 1–17.

Xiao, L. (2019). Asymptotic theory of penalized splines. *Electronic Journal of Statistics 13*(1), 747–794.

Yang, H., Z. Xu, I. King, and M. R. Lyu (2010). Online learning for group lasso. In J. Fürnkranz and T. Joachims (Eds.), *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 1191–1198. Omnipress.

Yao, F. and T. C. M. Lee (2008). On knot placement for penalized spline regression. *Journal of the Korean Statistical Society 37*(3), 259–267.

Yuan, Y. and S. Zhou (2012). Sequential B-spline surface construction using multiresolution data clouds. *Journal of Computing and Information Science in Engineering 12*, 021008.