Письма в ЭЧАЯ. 2008. Т. 5, № 3(145). С. 544–548



# DATA SMOOTHING BY SPLINES WITH FREE KNOTS

*N. D. Dikoussar*

Joint Institute for Nuclear Research, Dubna

*Cs. Török*

Technical University, Košice, Slovakia

A smoother based on an adaptive cubic model [1, 2] and splines with free knots is proposed. The model uses three reference data points and two parameters of control for estimation of a near optimal position of knots at the axis $x$ in autotracking mode. The data points are prethinned and corrected by local linear fitting. The coefficient table is obtained by standard spline procedure. The efficiency and the stability of the smoother with respect to random errors are shown on real noisy data.

Предложен алгоритм сглаживателя, основанный на адаптивной кубической модели [1, 2] и сплайнах со свободными узлами. Для вычисления близкой к оптимальной оценки положения узлов на оси абсцисс в режиме автоматического слежения модель использует три опорные точки и два управляющих параметра. Данные прореживаются и корректируются локальным линейным фитированием. Таблица коэффициентов получается с помощью стандартной сплайн-процедуры. Эффективность и устойчивость сглаживателя по отношению к случайным ошибкам показана на примерах обработки реальных данных с шумами.

PACS: 02.30.-f; 02.60.-x; 02.60.Gf

## INTRODUCTION

The paper proposes a smoothing procedure that produces a cubic spline $s(x; \mathbf{k}; \mathbf{c}_j)$, $j = \overline{1,k}$, with $k \geqslant 1$ internal knots from a set of data points

$$\{(x_i, \tilde{y}_i)\}_{i=1}^{n}, n \gg 4, \tag{1}$$

where $\tilde{y}_i = y_i + \epsilon_i$, $\epsilon_i \in N(0, \sigma^2)$; $\mathbf{k} = [x_1^*, x_2^*, \ldots, x_k^*]$, $x_j^* \in \{x_i\}_{i=1}^{n}$ is a set of knots detected automatically by the smoother, and $\mathbf{c}_j = [c_{0j}, c_{1j}, c_{2j}, c_{3j}]$ is a vector of coefficients of the model's polynomial at interval $[x_{j-1}^*, x_j^*]$, $j = \overline{1,k}$. The smooth function $s(x; \ldots)$ shows the association between $x_i$ and $\tilde{y}_i$ as follows:

$$\tilde{y}_i = s(x_i; \ldots) + r_i, \quad i = \overline{1,n}, \tag{2}$$

where $s(x_i; \ldots) = \hat{y}_i$ is the estimation of $\tilde{y}_i$, and the $r_i$ are residuals.

Smoothers have been used in many applications and are described in a number of references [3–5]. Recently, we have proposed a new «4-point approximation» based on the four-point transformation methodology [6–8]. The method and the algorithm (LOCUS-P) for approximation and smoothing data with no or moderate error have been described in [1,2,9].

The aim of this paper is to enhance the robustness of LOCUS-P for processing noisy data with complex dependency by piecewise cubic polynomials. There are several ways to solve this problem leveraging the autotracking piecewise cubic approximation. We mention two of them. The first one is that we can employ it to smoothed data. For smoothing (but not functionally describing) data, there are various methods, such as kernel smoothers [3] or Friedman's variable span smoother (*supersmoother*) [4]. As we provide for data description piecewise functions, it is not necessary to smooth every data point. It is sufficient to give local estimations for several data (trying not to lose any measurement) and to employ the autotracking piecewise approximation to the estimated data. The paper follows this second approach.

In [2] we studied approximation of data with complex dependence and no or moderate error, using a cubic model with a free parameter, in two stages: local and global approximation. The model plays a three-fold role: firstly, it is used on the local level for expressing the relation between $x$ and $y$, secondly, for the construction of an iterative scheme for the estimation of the model's parameter, and lastly, it enables a global continuous and smooth approximation in an automatic mode by piecewise cubic polynomials. While in the case of data with no or moderate errors the proposed autotracking piecewise approximation gives satisfactory results, in the case of errors with any variance (noisy data), there are problems with both the quality of the local approximants and the global smoothness (but not continuity). We succeeded in smoothing noisy data by autotracking piecewise approximation due to its combination with neural networks [10]. This paper proposes a solution without using NN.

Section 1 is a short introduction to the cubic model for piecewise approximation based on four points and provides the necessary formulas. The next section describes the way we correct and reduce the number of the reference points. Section 3 shows the results of smoothing real data.

## 1. A CUBIC MODEL FOR AUTOTRACKING PIECEWISE APPROXIMATION

Consider an additive model

$$\tilde{f} = f(x) + e. \tag{3}$$

We present the standard cubic polynomial $C = a_0 + a_1x + a_2x^2 + a_3x^3$ in the parametric form $s(\tau; \alpha, \beta, \mathbf{r}, \theta)$ with three fixed ($\mathbf{r}$), one free ($\theta$) and two control ($\alpha$ and $\beta$) parameters by Eq. (4). The curve $s(\tau; \alpha, \beta, \mathbf{r}, \theta)$ passes through four points $\{(x_*, f_*)\}, * = \tau, \alpha, \beta, 0$, where $\tau = x - x_0, \alpha = x_\alpha - x_0, \beta = x_\beta - x_0$, $f_* \equiv f(x_*)$. The vector $\mathbf{r} = [f_\alpha, f_\beta, f_0]^T$ is set up of the reference ordinates that are related to data points $\tilde{f}$. The abscissas $x_\tau, x_\alpha, x_\beta, x_0$ are used for evaluation of the vector of weight functions $\mathbf{w} = [w_1, w_2, w_3]^T$ defined by Eq. (5) and $Q$. $\theta$ is an unknown free parameter:

$$s(\tau; \alpha, \beta, \mathbf{r}, \theta) = \\ = \overbrace{f_\alpha w_1 + f_\beta w_2 + f_o w_3}^{\mathbf{w}^T\mathbf{r}} + \theta \overbrace{\tau(\tau - \alpha)(\tau - \beta)}^{Q} = \Pi(\tau; \alpha, \beta, \mathbf{r}) + \theta Q(\tau; \alpha, \beta), \tag{4}$$

where

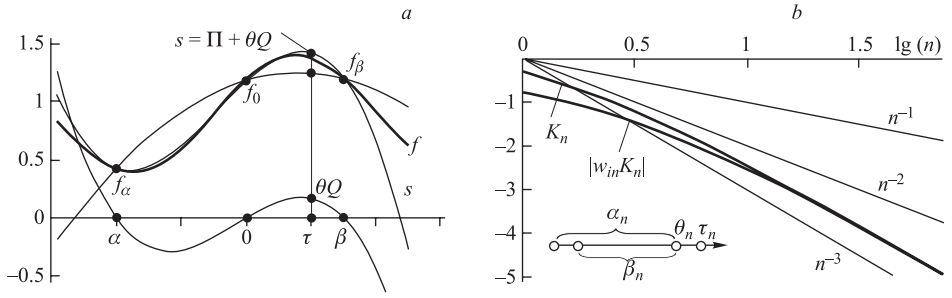


Fig. 1. The cubic model $s(\tau;\alpha,\beta,\mathbf{r},\theta)$ (*a*) and $|w_{in}K_n(\alpha_n,\beta_n)|$ in the log–log scale (*b*)

$$w_1=\frac{-\tau(\tau-\beta)}{\alpha\gamma},\quad w_2=\frac{\tau(\tau-\alpha)}{\beta\gamma},\quad w_3=\frac{(\tau-\alpha)(\tau-\beta)}{\alpha\beta},$$
$$\gamma=\beta-\alpha;\quad \alpha\beta\gamma\neq 0;\quad \sum_{i=1}^{3}w_i=1. \tag{5}$$

The quadratic parabola $\Pi(\tau;\alpha,\beta,\mathbf{r})=\mathbf{w}^T\mathbf{r}$ passes via three reference points and the cubic parabola $Q=\tau(\tau-\alpha)(\tau-\beta)$ is a «zeroing» parabola. Figure 1, *a* explains how the cubic parabola $s$ approximates a function $f$ using $\Pi$ and $\theta Q$ in the interval $[\alpha,\beta]$.

As the shape of curve Eq. (4) depends on the selection of the reference points $\mathbf{r}$, we can use the parameters $\alpha,\beta$ for controlling the error $e(x)=\tilde{f}(x)-s(x)$. For example, using the model (4) in *dynamic mode* we fix the points $(x_\alpha,\tilde{f}_\alpha)$ and $(x_\beta,\tilde{f}_\beta)$, and move the other two points $(x_0,\tilde{f}_0)$, $(x_\tau,\tilde{f}_\tau)$ with respect to the unmoved curve $\tilde{f}$. Minimization of $e^2(x)$ by the parameter $\theta$ leads to an iterative estimation of $\theta$:

$$\hat{\theta}_n=\hat{\theta}_{n-1}+K_n\overbrace{(\tilde{f}_n-\tilde{\Pi}_n-\hat{\theta}_{n-1}Q_n)}^{\varepsilon_n},\quad \hat{\theta}_0=0,\quad n=1,2,\ldots, \tag{6}$$

where $K_n=Q_n/\sum_{k=1}^{n}Q_k^2$ is an amplification factor and $\tilde{\Pi}_n=\Pi(\tau_n;\alpha_n,\beta_n,\tilde{\mathbf{r}}_n)$, $\tau_n=x_n-x_{0n}$, $\tilde{\mathbf{r}}=[\tilde{f}_{\alpha n},\tilde{f}_{\beta n},\tilde{f}_{0n},]^T$.

Equation (6) is a known adaptive procedure in which the output error is applied to input with the amplification factor $K_n(\alpha_n,\beta_n)$ that decreases as $\sim n^{-3}$, i.e., the errors $e_n$, $e_{\alpha n}$, $e_{\beta n}$ and $e_{0n}$ from Eq. (3) are suppressed near to a cubic-low because of $\mid w_{in}\mid\to 1$ for the above-described selection of $\alpha_n$ and $\beta_n$ [2] (Fig. 1, *b*).

For automatic tracking of a cubic segment of a curve the criterion of constancy of the third derivative of the cubic model is used [2].

## 2. CORRECTION OF THE REFERENCE ORDINATES

The critical part of the piecewise approximation by Eq. (6) in the case of noisy data is $\tilde{\Pi}_n=\Pi(\tau_n;\alpha_n,\beta_n,\tilde{\mathbf{r}}_n)$. It is clear that the reference ordinates $\tilde{\mathbf{r}}_n$ of the local approximants must be adjusted in some way. Here, we propose a process that does not need the correction of every data point.

Consider $M$ data points $(x_i, \tilde{f}_i)_{i=\overline{1,M}}$. We describe the algorithm of the piecewise functional smoothing in four steps with remarks:

1. Thinning of data points by selection of $N \ll M$ points. As we will see, the process can be applied to data with both equidistant and non-equidistant step.

2. Local estimation of the ordinates $\hat{f}$ of the selected $N$ points. There are many ways how to get good local point estimations. They have to be effective and take into account every $M$ data points.

3. Reduction of the estimated $N$ selected points using Eqs. (4)–(6) to $K$ points. The detection of $K$ knots by the first stage of the autotracking piecewise cubic approximation is executed on the $N$ estimated points, so the reference ordinates in Eq. (6) have been corrected.

4. Construction of integral approximants based on the reduced $K$ number of estimated points. To get continuous integral estimation the methods and formulas from the second stage of the autotracking piecewise cubic approximation can be leveraged, see [1, 2, 9]. To get not only continuous approximants, but also approximants with continuous first and second derivatives, the spline table can be computed based on the reduced $K$ estimated points from the third step.

## 3. EXAMPLES

In the previous section we described shortly in four steps the smoothing process based on local estimation (step 2) and the autotracking piecewise cubic approximation (step 3). To demonstrate the process we considered real noisy data with both equidistant and non-equidistant step. From the three data sets the first one shows the most complex relation. The figures contain the original data of length $M$ denoted by little squares, the continuous spline smoothers, the residuals at the bottom of the pictures, the histograms constructed from the residuals, and the verticals denoting the endpoints of $K$ segments. We also provide the number $N$ of selected and estimated data points.

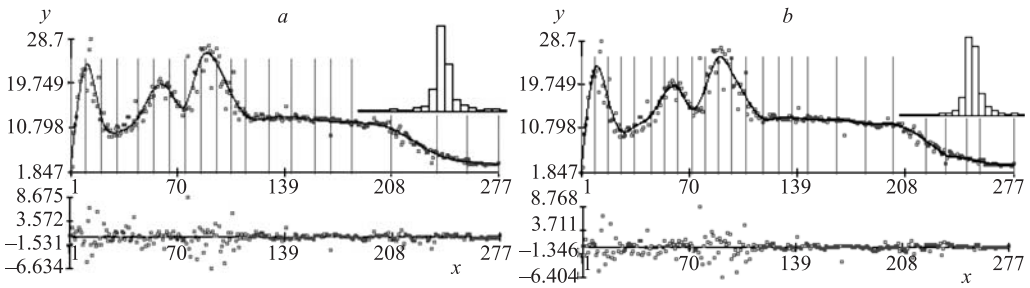


Fig. 2. Cross section for $\pi^- p$ collision: *a*) $M = 277$, $N = 130$, $K = 24$; *b*) $M = 277$, $N = 57$, $K = 20$

Figure 2 illustrates smoothing data with equidistant step, the cross sections for $\pi^- p$ collision [11]. Although *a* and *b* splines were evaluated based roughly on every second and fifth locally corrected data, thanks to the autotracking knot detection from the third step their number was reduced approximately five and three times, to 24 and 20, respectively.

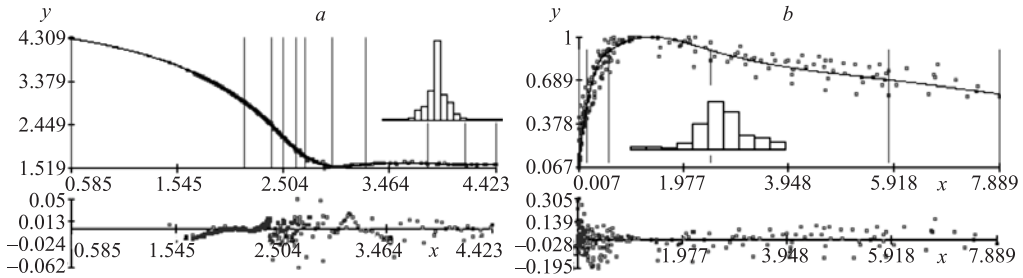


Fig. 3. Data with non-equidistant step: cross section for $np$ collision (*a*) and concrete characteristic (*b*): *a*) $M = 325$, $N = 40$, $K = 10$; *b*) $M = 196$, $N = 18$, $K = 5$

In Fig. 3 we give the smoothing results of two data sets with non-equidistant step: *a* illustrates the cross sections for $np$ collision and *b* — the resistance ratio. The reduced autotracked number of segments and the quality of the piecewise approximants are adequate and acceptable in all cases.

## CONCLUSIONS

The paper describes a smoothing process with local estimations and automatic knot detection for describing noisy data with complex dependence by piecewise continuous cubic polynomials. The resulting spline tables are slim and the splines provide for both simulated and real noisy data satisfactory approximation.

**Acknowledgements.** We thank L. S. Azhgirej, A. Polansky, Ďuricová & Rovňák for calling our attention to and providing the data. This work was partially supported by VEGA Grant 1/1006/04 of MŠ SR.

REFERENCES

1. *Dikoussar N. D.* JINR Commun. P10-99-168. Dubna, 1999 (in Russian);
   *Dikoussar N. D.* JINR Preprint E10-2001-48. Dubna, 2001.
2. *Dikoussar N. D., Török Cs.* // Math. Model. 2006. V. 18, No. 3. P. 23–40 (in Russian).
3. *Härdle W.* Applied Nonparametric Regression. Cambridge Univ. Press, 1990.
4. *Friedman J.* A Variable Span Smoother. SLAC PUB-3477. Stanford, 1984.
5. *Silverman B. W.* // J. Am. Statist. Assn. 1984. V. 19.
6. *Dikoussar N. D.* // Math. Model. 1991. V. 10, No. 3. P. 50–64 (in Russian).
7. *Dikoussar N. D.* // Comp. Phys. Commun. 1994. V. 79. P. 39–51.
8. *Dikoussar N. D.* // Comp. Phys. Commun. 1997. V. 99. P. 235–254.
9. *Török Cs., Dikoussar N. D.* JINR Commun. P10-2004-202. Dubna, 2004.
10. *Révayová M., Török Cs.* // Kibernetika. 2006 (submitted).
11. Eur. Phys. J. C. Review of Particle Physics. Springer, 2000. P. 235.