

Wavelet smoothing for curves in more than one dimension

Davide Pigoli and Laura M. Sangalli

Abstract The estimation of smooth functions from their noisy and discrete observation is the first step in Functional Data Analysis. The choice of the basis functions is crucial in the process, since its properties influence the subsequent analysis. Wavelets offer functional basis systems with the property of localization both in space and frequency. We developed a procedure to use wavelet bases for the estimation of curves in more than one dimension, with a particular focus on the possibility of obtaining also estimates of curves derivatives from wavelet expansion. This method is tested on the estimation of centreline and radius of the Internal Carotid Artery (ICA) for patients suspected to be affected by cerebral aneurism.

Key words: Wavelets, Functional data analysis, Smoothing

1 Introduction

Functional Data Analysis (FDA) is the branch of statistics which focuses on data that can be seen as the observed value of a functional random variable. However, from a practical point of view, every data is observed on a discrete grid and a measurement error is also present. A crucial step of the analysis thus consists in the estimation of the continuous functional data starting from its discrete observation. In this process basis functions play an essential role. Usual choices are Fourier basis functions and spline basis functions (see [8]). Wavelet basis systems have been mainly used for problems in which there is no interest in derivatives, because of the absence of

Davide Pigoli
MOX, Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133
Milano, Italy, e-mail: davide.pigoli@mail.polimi.it

Laura M. Sangalli
MOX, Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133
Milano, Italy, e-mail: laura.sangalli@polimi.it

close analytical form for smooth wavelets. This has restricted their application to a small part of the FDA field. To overcome this limitation, we look for a way to obtain derivatives of wavelet estimated data. Then, we extend traditional wavelet estimators to curves in general dimensions; this requires the development of a new estimation procedure which takes into account simultaneously all the coordinates of the multidimensional curve.

2 An overview on wavelets

In this section we briefly introduce wavelet basis systems for $L^2(\mathbb{R})$. Wavelets are defined starting from an orthogonal multiresolution:

Definition 1. Let $\{V_j\}$ be a sequence of closed subspaces $V_j \subseteq L^2(\mathbb{R})$ and $\varphi \in V_0$. An orthogonal multiresolution for $L^2(\mathbb{R})$ is a couple $(\{V_j\}, \varphi)$ so that:

1. $V_j \subset V_{j+1}$
2. $\bigcup_j V_j = L^2(\mathbb{R})$ and $\bigcap_{j=-\infty}^+ V_j = \{0\}$
3. $t \mapsto f(t) \in V_j \Leftrightarrow t \mapsto f(2t) \in V_{j+1}$
4. $\{\varphi(t-k)\}_k$ is an orthonormal basis for V_0 and $\int_{\mathbb{R}} \varphi \neq 0$

The projections of $f \in L^2(\mathbb{R})$ on the sequence $\{V_j\}_{j \in \mathbb{Z}}$ give a progressively better approximation of f as j increases. The function φ is called *scaling function* or *father wavelet*. Thanks to property 3 above, $\{2^{j/2}\varphi(2^j t - k)\}_k$ is an orthonormal basis for V_j . However, it is often more useful exploring the detail information needed to go from the space V_j to the space V_{j+1} , starting from a *coarse* space V_0 . This is the reason to introduce the sequence of the complement spaces $W_j = V_{j+1} \setminus V_j$. A *mother wavelet* is a function $\psi \in W_0$ so that $\{\psi(t-m)\}_{m \in \mathbb{Z}}$ is a basis for W_0 . Since

$$L^2(\mathbb{R}) = \bigoplus_{j \in \mathbb{Z}} W_j, \quad (1)$$

$\{\psi_{j,k}(t) = 2^{j/2}\psi(2^j t - k)\}_{j,k}$ is an orthonormal basis for $L^2(\mathbb{R})$. Therefore, for each $f \in L^2(\mathbb{R})$, we have

$$\begin{aligned} f &= \sum_j \sum_k \langle f, \psi_{j,k} \rangle \psi_{j,k} = \sum_k \langle f, \varphi_{j_0,k} \rangle \varphi_{j_0,k} + \sum_{j=j_0}^{+\infty} \sum_k \langle f, \psi_{j,k} \rangle \psi_{j,k} = \\ &= \sum_k s_{j_0,k} \varphi_{j_0,k} + \sum_{j=j_0}^{+\infty} \sum_k d_{j,k} \psi_{j,k} \end{aligned} \quad (2)$$

where $\langle \cdot, \cdot \rangle$ indicates the scalar product in $L^2(\mathbb{R})$. The coefficients $\{s_{j_0,k}\}_{k \in \mathbb{Z}}$, $\{d_{j,k}\}_{j \in \mathbb{Z} \cap \{j > j_0\}, k \in \mathbb{Z}}$ are called *discrete wavelet transform* (DWT) of f . It can be shown that φ and ψ satisfy the dilation/refinement equations:

$$\varphi(t) = \sum_k h_k \sqrt{2} \varphi(2t - k)$$

$$\psi(t) = \sum_k g_k \sqrt{2} \varphi(2t - k)$$

for some sequences h_k and g_k , named respectively scaling filter and wavelet filter. These equations are essential for the development of the Fast Wavelet Transform, which computes the DWT in $O(n)$ operations. Moreover, smooth and compactly supported wavelets have no analytical form and so they are defined by their scaling and wavelet filters. For a systematic introduction to wavelets, see e.g. [6].

3 Wavelet smoothing of functional data

As we have seen above, every function $f \in L^2(\mathbb{R})$ can be represented by a scaling / wavelet basis system. Now we would like to use this fact to build an estimator of the functional data f starting from its discrete observation. The statistical model is

$$x_i = f(s_i) + \varepsilon_i, \quad i = 1, \dots, N, \quad N = 2^J, \quad J \in \mathbb{N} \quad (3)$$

where $s_i = \frac{i}{N}$ are equally spaced points and $\varepsilon_i \sim N(0, \sigma^2)$ are independent and identically distributed errors. The wavelet smoothing procedure consists in changing over to wavelet domain, where the model becomes

$$d_{j,k} = d_{j,k}^0 + \rho_{j,k},$$

with $d_{j,k}^0$ being the true wavelet coefficients of f and $\rho_{j,k} \sim N(0, \sigma_d^2)$ the wavelet transform of the error. A first idea to estimate $d_{j,k}^0$, starting from the empirical coefficients $d_{j,k}$, consists in fixing a threshold t and considering all coefficients below this threshold as coming only from noise, thus setting $\hat{d}_{j,k} = d_{j,k} \mathbb{I}_{\{|d_{j,k}| \leq t\}}$. A more refined idea consists in also shrinking the coefficients above this threshold, with the aim to removing the component due to noise. In particular, the wavelet estimator becomes

$$\hat{d}_{j,k} = \text{sign}(d_{j,k}) (|d_{j,k}| - t)_+.$$

and the corresponding estimated functional data is

$$\hat{f}(s) = \sum_{k=0}^{2^{j_0}-1} s_{j_0,k} \varphi_{j_0,k}(s) + \sum_{j=j_0}^{J-1} \sum_{k=0}^{2^j-1} \hat{d}_{j,k} \psi_{j,k}(s). \quad (4)$$

The problem is of course the choice of the threshold t . Many strategies have been proposed (see e.g. [2], [4] or [5]), among which a popular one is the so called *universal threshold*:

$$t = \hat{\sigma}_d \sqrt{2 \log N}$$

As mentioned in the Introduction, the application of wavelet basis systems for fitting functional data has been mainly limited to problems in which derivatives were not required, due to the absence of a close analytical form for wavelets smooth enough for this purpose. Anyway, φ , ψ and their derivatives can be computed on the dyadic points starting from the dilation equation. In [7] we exploit this fact to obtain curve derivatives starting from wavelet expansion.

4 Wavelet estimation for curves in more than one dimension

Let us now extend wavelet based estimation techniques to the case of curves in more than one dimension. The function $\mathbf{f}$ we want to estimate has thus the form

$$\mathbf{f}: \mathbb{R} \ni s \mapsto (f_1, \dots, f_p) \in \mathbb{R}^p$$

which describes parametric curves in p dimensions. The observed values are generated by the model

$$\mathbf{x}_i = \mathbf{f}(s_i) + \varepsilon_i \quad i = 1, \dots, N = 2^J \quad (5)$$

where ε_i is a multinormal error with zero mean and variance and covariance matrix $\sigma^2 \mathbb{I}$. Our aim is to estimate the function $\mathbf{f}$ and its derivatives. A first idea could be to estimate each coordinate function $f_1, \dots, f_p$ independently, using well-known wavelet techniques for the denoising of 1D signals. However, if the curve has a significant feature in some points of the physical space, we expect that this will be reflected on all p coordinates concurrently. Thus, we develop a thresholding procedure taking into account the vectorial structure of the function we want to estimate. In particular, we would like our thresholding procedure to estimate as zero the wavelet coefficient of a coordinate function only if this coefficient is estimate as zero for all coordinates. In this way, the same wavelet basis functions are used for the estimation of $f_1, \dots, f_p$. Starting from the model (5) and using the orthogonality of the wavelet transform, we know that

$$\mathbf{d}_{j,k} = \mathbf{d}_{j,k}^0 + \rho_{j,k},$$

where $\mathbf{d}_{j,k}^0 \in \mathbb{R}^p$ is the vector of the true wavelet coefficients of $\mathbf{f}$, while $\rho_{j,k}$ is the wavelet transform of the noise and it has a multivariate normal distribution with covariance matrix $\sigma_d^2 \mathbb{I}$. In order to decide if $\mathbf{d}_{j,k}^0$ can be estimated as the null vector, we focus on $\|\mathbf{d}_{j,k}\|_2^2$. In order to fix a threshold for this quantity, aiming to contain all coefficients coming from noise, we exploit the fact that $\|\frac{\rho_{j,k}}{\sigma_d}\|_2^2 \sim \chi^2(p)$. Therefore, we look for a thresholding value which contains with high probability n draws from a random variable with $\chi^2(p)$ distribution. In [7] we show that a good choice for this threshold is $c_p \log n$, with $c_1 = 2$ and $c_p = 3$ for $p \geq 2$. Fig. 1 shows some simulations in support of this result. It can be noticed that, for $p \geq 2$, the convergence rate decreases as p increases. This makes sense, because we use the same threshold to control the maximum of random variables with increasing mean and variance.

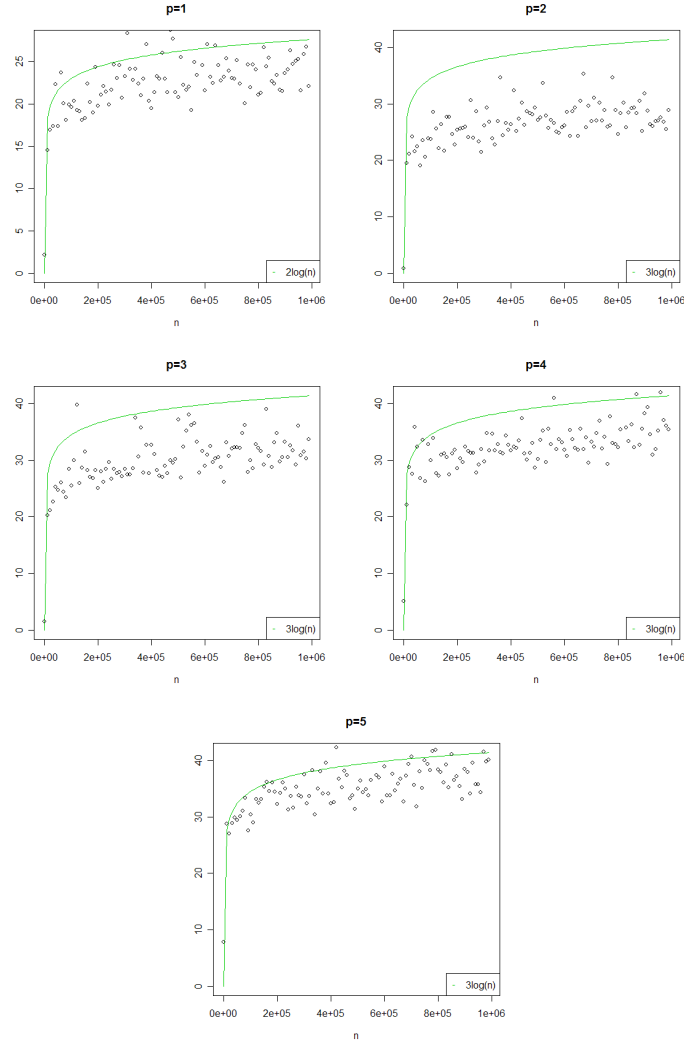


Fig. 1 Maximum of n draws from a $\chi^2(p)$ distribution plotted vs n . The green line indicates the threshold t_p .

Thus we shall use the following threshold on $\|\mathbf{d}_{j,k}\|_2^2$

$$t_p = \hat{\sigma}_d^2(c_p \log N).$$

This threshold is a generalization of the *universal threshold* proposed in [5] for the 1D case. The empirical coefficients are vectors of $\mathbb{R}^p$ and $\sqrt{t_p}$ is a threshold on their euclidian norm. Hence we consequently apply the shrinkage on their norm,

proposing the following wavelet estimator for curves in p dimensions

$$\hat{\mathbf{d}}_{j,k} = \left(1 - \frac{\sqrt{t_p}}{\|\mathbf{d}_{j,k}\|_2}\right)_+ \mathbf{d}_{j,k}$$

(see [7]). In this way, estimated wavelet coefficients have the same direction of empirical ones, as vectors in $\mathbb{R}^p$. This is based on the hypothesis that the variance of the error is equal in the p directions. Thus, the direction of the vectors $\mathbf{d}_{j,k}$ can be considered as mainly determined by the true coefficients $\mathbf{d}_{j,k}^0$.

5 An application to the geometry of the Inner Carotid Artery (ICA)

In this section we use wavelets for the estimation of radius and curvature of the Inner Carotid Artery (ICA) for 65 patients suspected to be affected by a cerebral aneurism, treated at Niguarda Ca'Granda Hospital (Milan). Our data are the discrete observations of the radius profile of the vessel and of the three space coordinates of the vessel centreline, supplied by the Vascular Modeling ToolKit (see [1]). The analyzed data belong to the AneuRisk Project¹, a research programme that investigated the role of vascular geometry in the pathogenesis of cerebral aneurisms.

The data are of course affected by measurement and reconstruction errors. In [9] estimates of the three-dimensional ICA centrelines and their derivatives, and thus of centreline curvature, have been obtained via free-knot regression splines. Here, we want to estimate ICAs geometrical features, in particular vessel curvature and vessel radius, through wavelets. We shall use in particular Daubechies “extremal phase” wavelets with 10 vanishing moment (see [3]). The importance of vessel curvature and vessel radius in the study of aneurismal pathology is discussed for instance in [9] and [10]. Fig. 3, left panel, gives a 3D visualization of the estimated vessel centreline for patient 1. The same figure, right panel, shows the estimated curvature (obtained from the estimates of first and second derivatives) for the vessel centreline of all 65 patients of the AneuRisk Dataset. Fig. 2 displays the estimates of the radius of the artery and its derivatives for the 65 patients. Fig. 4 shows the boxplots of the Average Square Error (ASE) of the estimates of the 65 centerlines obtained by wavelets and by free-knot regression splines, and the boxplots of the Average Square Error of the estimates of the centerlines first derivatives with respect to first finite differences (ASEdiff). A simultaneous comparison of these boxplots shows that the two methods provide comparable results on these data (free-knot regression splines

¹ The AneuRisk project is a joint research programme involving the MOX Laboratory for Modeling and Scientific Computing (Dipartimento di Matematica, Politecnico di Milano), the Laboratory of Biological Structure Mechanics (Dipartimento di Ingegneria Strutturale, Politecnico di Milano), Istituto Mario Negri (Ranica), Ospedale Niguarda Ca' Granda (Milano) and Ospedale Maggiore Policlinico (Milano). The project has been supported by the Fondazione Politecnico di Milano and Siemens Medical Solutions Italia.

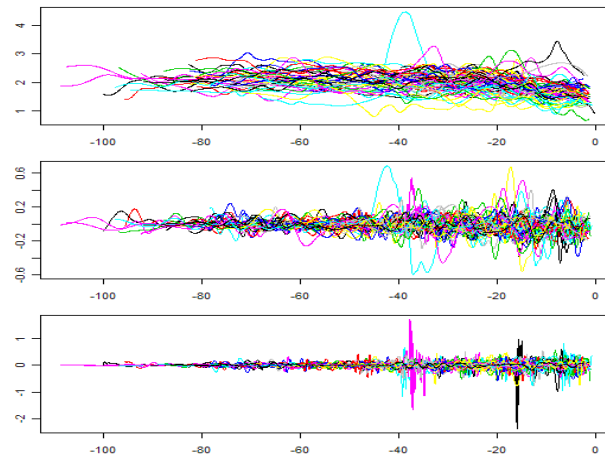


Fig. 2 Estimated radius (top) and its first (center) and second (bottom) derivatives for the 65 patients.

having the advantage of dealing with general grids of points, not necessarily of 2^J evenly-spaced points). More extensive comparisons between wavelets and various smoothing methods will be published elsewhere.

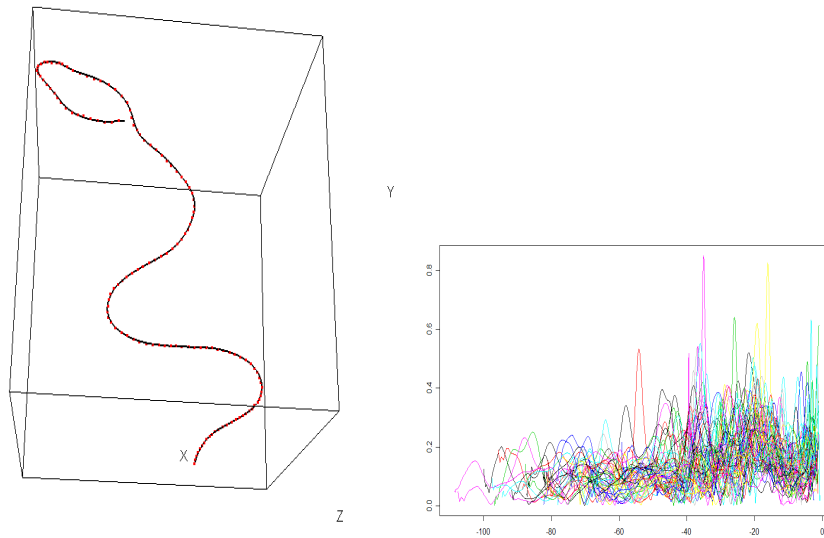


Fig. 3 Left: Estimated centreline for patient 1 (the green bullets indicate the estimated points of approximately zero curvature). Right: Estimated curvature for the 65 patients.

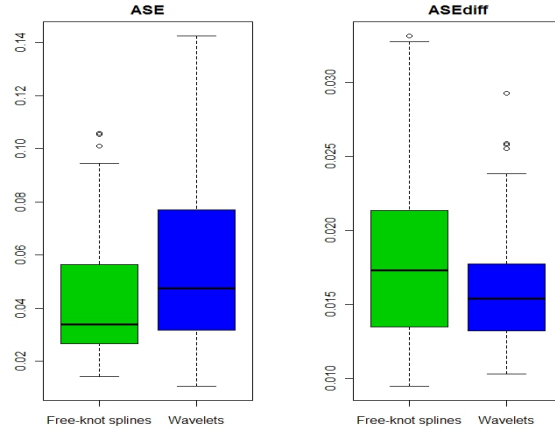


Fig. 4 Comparison between free knots regression splines and wavelets, based on $ASE = \frac{1}{n} \sum_{k=1}^n [(x(s_k) - \hat{x}(s_k))^2 + (y(s_k) - \hat{y}(s_k))^2 + (z(s_k) - \hat{z}(s_k))^2]$ and $ASEdiff = \frac{1}{n} \sum_{k=1}^n [(Dx(s_k) - \hat{x}'(s_k))^2 + (Dy(s_k) - \hat{y}'(s_k))^2 + (Dz(s_k) - \hat{z}'(s_k))^2]$ where Dx, Dy and Dz are the first central finite differences for x, y , and z .

Acknowledgements. We are grateful to Professor Piercesare Secchi for his helpful suggestions and comments. We would also like to thank Dr. Marco Verani for useful and constructive discussions related to this work.

References

1. Antiga, L., Ene-Iordache, B., Remuzzi, A.: Computational geometry for patient-specific reconstruction and meshing of blood vessels from MR and CT angiography. *IEEE Trans. Med. Imaging* **22**, 674–684 (2003)
2. Cai, T.T., Zhou H.H.: A data-driven block thresholding approach to wavelet estimation. *Ann. Stat.* **37**, 569–595 (2009)
3. Daubechies, I.: Orthonormal basis of compactly supported wavelets. *Commun. Pur. Appl. Math.* **41**, 909–996 (1988)
4. Donoho, D.L., Johnstone, I. M.: Adapting to Unknown Smoothness via Wavelet Shrinkage, *J. Am. Stat. Assoc.* **90**, 1200–1224 (1995)
5. Donoho, D.L., Johnstone, I. M., Kerkyacharian, G., Picard, D.: Wavelet Shrinkage: Asymptopia. *J. Roy. Stat. Soc. B Met.* **57**, 301–369 (1995)
6. Nason, G.P.: *Wavelet Methods in Statistics with R*, Springer, New-York (2008)
7. Pigoli, D., Sangalli, L.M.: *Wavelets in Functional Data Analysis: estimation of multidimensional curves with their derivatives*, work in progress
8. Ramsay, J.O., Silverman, B.W.: *Functional data analysis*. Second edition. Springer, New York (2005).
9. Sangalli, L.M., Secchi, P., Vantini S., Veneziani, A.: Efficient estimation of three-dimensional curves and their derivatives by free-knot regression splines, applied to the analysis of inner carotid artery centrelines. *J. Roy. Stat. Soc. C-App.* **58**, 285–306 (2009)
10. Sangalli, L.M., Secchi, P., Vantini S., Veneziani, A.: A Case Study in Exploratory Functional Data Analysis: Geometrical Features of the Internal Carotid Artery. *J. Am. Stat. Assoc.* **104**, 37–48 (2009)