Nonparametric Derivative Estimation

Kris De Brabanter [a] Jos De Brabanter [a b] Bart De Moor [a]

[a] *Department of Electrical Engineering (ESAT-SCD), Katholieke Universiteit Leuven, Kasteelpark Arenberg 10 B-3001 Leuven, Belgium*

[b] *Hogeschool KaHo Sint-Lieven (Associatie K.U.Leuven), Departement Industrieel Ingenieur, G. Desmetstraat 1, B-9000 Gent, Belgium*

Abstract

We present a simple but effective fully automated framework for estimating first order derivatives nonparametrically. Derivative estimation plays an important role in the exploration of structures in curves (jump detection and discontinuities), comparison of regression curves, analysis of human growth data, etc. Hence, the study of estimating derivatives nonparametrically is equally important as regression estimation. Via empirical first order derivatives we approximate the first order derivative and create a new data set which can be smoothed by any nonparametric regression estimator. However, the new data sets created by this technique are not independent and identically distributed (i.i.d.) random variables anymore. As a consequence, automated model selection criteria (data-driven procedures) break down. Therefore, we modify the model selection criterion so it can handle this dependency (correlation) without requiring any prior knowledge about its structure.

keywords: nonparametric derivative estimation, model selection, empirical first order derivatives

1 Introduction

Ever since the introduction of nonparametric estimators for density estimation, regression, etc. in the mid 1950s and early 1960s, their popularity has increased over the years. Mainly, this is due to the fact that statisticians realized that pure parametric thinking in curve estimations often does not meet the need for flexibility in data analysis. Many of their properties have been rigorously investigated and are well understood, see e.g [4, 11]. Although the importance of regression estimation is indisputable, sometimes the derivative of the regression estimate can be equally important. This is the case in the exploration of structures in curves [3] (jump detection and discontinuities), inference of significant features in data, trend analysis in time series [9], comparison of regression curves [7], analysis of human growth data [8], the characterization of submicroscopic nanoparticles from scattering data [2] and inferring chemical compositions. All the previous analysis techniques are based on the inference about slopes (and hence the derivative) of the regression estimates. Therefore, the study of estimating derivatives (first and higher orders) nonparametrically is equally important as regression estimation.

Consider the bivariate data $(x_1,Y_1),\ldots,(x_n,Y_n)$ which form an independent and identically distributed (i.i.d) sample from a population (x,Y) where $x \in \mathbb{R}$ and $Y \in \mathbb{R}$. Denote by $m(x) = \mathbf{E}[Y]$ the regression function. The data is regarded to be generated from the model

$$Y = m(x) + e, \tag{1}$$

where $\mathbf{E}[e] = 0$, $\mathbf{Var}[e] = \sigma^2 < \infty$ and x and e are independent. The aim of this paper is to estimate the derivative m' of the regression function m.

This paper is organized as follows. Our nonparametric estimator of choice is illustrated in Section 2. The proposed method for estimating first order derivatives nonparametrically and model selection issues are discussed in Section 3. Simulations are presented in Section 4. Conclusions are given in Section 5.

2 Least squares support vector machines for regression

Given a training set defined as $\mathcal{D}_n = \{(x_k,Y_k) : x_k \in \mathbb{R}^d, Y_k \in \mathbb{R}; k = 1,\ldots,n\}$. Then least squares support vector machines for regression are formulated as follows [10]

$$\begin{aligned} \min_{w,b,e} \mathcal{J}(w,e) &= \tfrac{1}{2}w^T w + \tfrac{\gamma}{2}\sum_{k=1}^{n} e_k^2 \\ \text{s.t.} \quad Y_k &= w^T\varphi(x_k) + b + e_k, \quad k = 1,\ldots,n, \end{aligned} \tag{2}$$

where $e_k \in \mathbb{R}$ are assumed to be i.i.d. random errors with zero mean and finite variance, $\varphi : \mathbb{R}^d \to \mathbb{R}^{n_h}$ is the feature map to the high dimensional feature space (possibly infinite dimensional) and $w \in \mathbb{R}^{n_h}$, $b \in \mathbb{R}$. Note that this cost function consists of a residual sum of squares fitting error and a regularization term, which is also a standard procedure for the training of Multi-Layer Perceptrons (MLPs). Also, the above formulation is related to ridge regression.

To solve the optimization problem (2) in the dual space, one defines the Lagrangian

$$\mathfrak{L}(w,b,e;\alpha) = \frac{1}{2}w^T w + \frac{\gamma}{2}\sum_{i=1}^{n} e_i^2 - \sum_{i=1}^{n}\alpha_i\{w^T\varphi(x_i) + b + e_i - Y_i\},$$

with Lagrange multipliers $\alpha_i \in \mathbb{R}$ (called support vectors). The conditions for optimality are given by

$$\begin{cases} \dfrac{\partial\mathfrak{L}}{\partial w} = 0 \quad \to \quad w = \sum_{i=1}^{n}\alpha_i\varphi(x_i) \\ \dfrac{\partial\mathfrak{L}}{\partial b} = 0 \quad \to \quad \sum_{i=1}^{n}\alpha_i = 0 \\ \dfrac{\partial\mathfrak{L}}{\partial e_i} = 0 \quad \to \quad \alpha_i = \gamma e_i, \qquad\qquad i = 1,\ldots,n \\ \dfrac{\partial\mathfrak{L}}{\partial \alpha_i} = 0 \quad \to \quad w^T\varphi(x_i) + b + e_i - Y_i = 0, \quad i = 1,\ldots,n. \end{cases}$$

After elimination of w and e the solution yields

$$\left[\begin{array}{c|c} 0 & 1_n^T \\ \hline 1_n & \Omega + \frac{1}{\gamma}I_n \end{array}\right]\left[\begin{array}{c} b \\ \hline \alpha \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline Y \end{array}\right],$$

with $Y = (Y_1,\ldots,Y_n)^T$, $1_n = (1,\ldots,1)^T$, $\alpha = (\alpha_1,\ldots,\alpha_n)^T$ and $\Omega_{kl} = \varphi(x_k)^T\varphi(x_l) = K(x_k,x_l)$, with $K(x_k,x_l)$ positive definite, for $k,l = 1,\ldots,n$. According to Mercer's theorem, the resulting LS-SVM model for function estimation becomes

$$\hat{m}(x) = \sum_{k=1}^{n}\hat{\alpha}_k K(x,x_k) + \hat{b}. \tag{3}$$

In this paper we take $K(x_i,x_j) = (2\pi)^{-d/2}\exp\left(-\frac{\|x_i - x_j\|_2^2}{2h^2}\right)$ (Gaussian kernel).

3 Derivative estimation

In this section we first illustrate the principle of empirical first order derivatives and how they can be used together with a nonparametric regression estimator to estimate first order derivatives. As the created data sets by this method are no longer i.i.d. random variables, data-driven model selection procedures will break down. Second, we illustrate how to handle dependent data in the data-driven procedures.

3.1 Empirical first order derivatives

Given the nonparametric regression estimate (3), it would be tempting to differentiate it w.r.t. the independent variable. Such a procedure can only work well if the original regression function is extremely well estimated. Otherwise, it can lead to wrong derivative estimates when the data is noisy. This is due to the

fact that we already make an error (maybe small) when estimating the regression function. Differentiating this estimate will only result in an accumulation of errors which increases with the order of the derivative. A possible solution to avoid this problem is by using the first order difference quotient

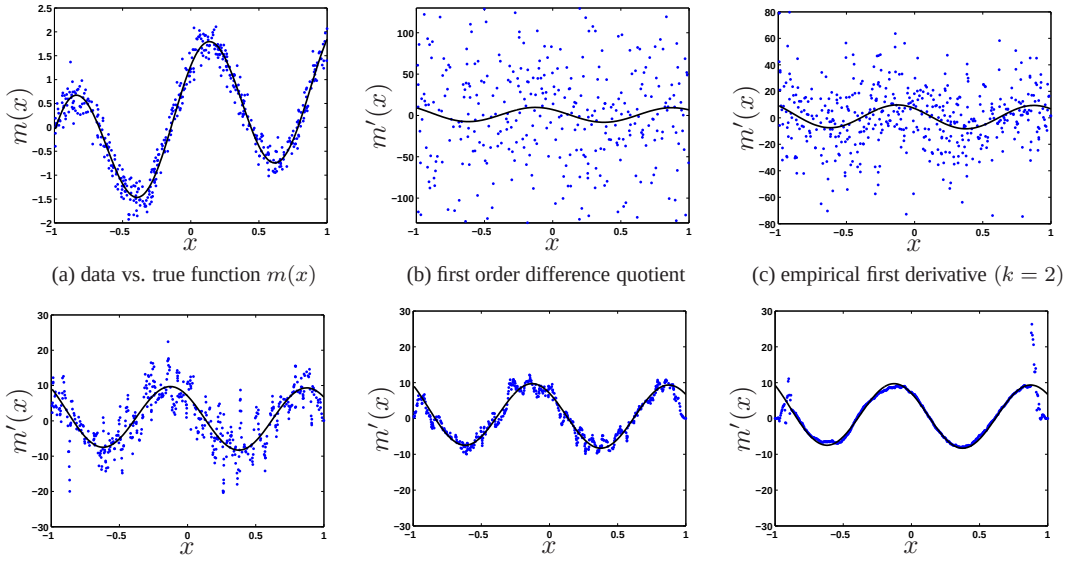
$$Y_i^{(1)} = \frac{Y_i - Y_{i-1}}{x_i - x_{i-1}}$$

as a noise corrupted version of $m'(x_i)$ where the superscript (1) signifies that $Y_i^{(1)}$ is a noise corrupted version of the first (true) derivative. Such an approach has been used by [5] to estimate derivatives nonparametrically. Although this seems again intuitively the right way, the generated new data will be very noisy and as a result it will be difficult to estimate the derivative function. A better way to generate the raw data for derivative estimation is to use a variance-reducing linear combination of symmetric (about i) difference quotients

$$Y_i^{(1)} = \sum_{j=1}^{k} w_j \cdot \left(\frac{Y_{i+j} - Y_{i-j}}{x_{i+j} - x_{i-j}} \right), \tag{4}$$

where $k \in \mathbb{N}\backslash\{0\}$ and the weights $w_1,\ldots,w_k$ sum up to one. The linear combination (4) is valid for $k+1 \leq i \leq n-k$. For $2 \leq i \leq k$ or $n-k+1 \leq i \leq n-1$ we define $Y_i^{(1)}$ by replacing $\sum_{j=1}^{k}$ in (4) by $\sum_{j=1}^{k(i)}$ where $k(i) = \min\{i-1, n-i\}$ and replacing $w_1,\ldots,w_k$ by $w_1/\sum_{j=1}^{k(i)} w_j,\ldots,w_{k(i)}/\sum_{j=1}^{k(i)} w_j$. finally, for $i = 1$ and $i = n$ we define $Y_1^{(1)}$ and $Y_n^{(1)}$ to coincide with $Y_2^{(1)}$ and $Y_{n-1}^{(1)}$. The proportion of indices i falling between $k+1$ and $n-k$ approaches 1 as n increases, so this boundary issue becomes smaller as n becomes larger. Alternatively, one may just leave $Y_i^{(1)}$ undefined for indices i not falling between $k+1$ and $n-k$. In this paper we will use the first principle to estimate the derivatives.

Linear combinations as in (4) are frequently used in finite element theory and are useful in the numerical solution of differential equations [6]. However, the weights used for solving differential equations are not appropriate here because of the random errors in model (1). By choosing the weights $w_j = j^2/\sum_{l=1}^{k} l^2$ with $j \in \{1,\ldots,k\}$ the variance of $Y_i^{(1)}$, for $k+1 \leq i \leq n-k$, is minimized [2]. Since we consider the data to be one dimensional we can visualize each generated data set for different values of k. The optimal value of k can be found for example via leave-one-out cross-validation (LOO-CV). See the next paragraph for more details. Figure 1 displays the empirical first derivative for $k \in \{2,6,12,25\}$ generated from model (1) with $m(x) = \sin(2\pi x) + \cos(2\pi x) + \log(4/3 + x)$, $n = 500$ equispaced points and $e \sim \mathcal{N}(0,0.1^2)$.



(a) data vs. true function $m(x)$ (b) first order difference quotient (c) empirical first derivative $(k = 2)$

(d) empirical first derivative $(k = 6)$ (e) empirical first derivative $(k = 12)$ (f) empirical first derivative $(k = 25)$

Figure 1: (a) Simulated data set of size $n = 500$ equispaced points from model (1) with $m(x) = \sin(2\pi x) + \cos(2\pi x) + \log(4/3 + x)$ and $e \sim \mathcal{N}(0,0.1^2)$; (b) first order difference quotients which are barely distinguishable from noise. As a reference, the true derivative is also displayed (full line); (c)-(f) empirical first derivatives for $k \in \{2,6,12,25\}$.

3.2 Model selection

By using the previous technique, we have created a new data set i.e. $(x_1,Y_1^{(1)}),\ldots,(x_n,Y_n^{(1)})$. This new data set can now be smoothed to obtain an estimate of the first order derivative of the regression function. However, since each $Y_i^{(1)}$ is formed as a sum of differences of consecutive Y (see (4)), the $Y_i^{(1)}$ $i=1\ldots,n$ are not independent anymore. As a consequence, all model selection criteria cannot be legitimately be applied anymore since they are based on model assumption (1). We briefly illustrated how to modify the leave-one-out cross-validation procedure. This summary is based on [1].

Since our smoother of choice is LS-SVM we require a positive definite kernel. Therefore we require a two-step model selection criteria (see [1]). Consider the Nadaraya-Watson (NW) kernel smoother defined as

$$\hat{m}(x)=\sum_{i=1}^{n}\frac{K(\frac{x-x_i}{h})Y_i^{(1)}}{\sum_{j=1}^{n}K(\frac{x-x_j}{h})},$$

where h is the bandwidth of the kernel K. An optimal bandwidth h can for example be found by minimizing the leave-one-out cross-validation (LOO-CV) score function

$$\text{LOO-CV}(h)=\frac{1}{n}\sum_{i=1}^{n}\left(Y_i^{(1)}-\hat{m}^{(-i)}(x_i;h)\right)^2, \tag{5}$$

where $\hat{m}^{(-i)}(x_i;h)$ denotes the leave-one-out estimator where point i is left out from the training. For notational ease, the dependence on the bandwidth h will be suppressed. Then, according to [1] (see Theorem 3), by taking a kernel function satisfying $K(0)=0$ removes the correlation structure without requiring any prior knowledge about its structure. Denote a kernel that is satisfying $K(0)=0$ by $\tilde{K}$. In this paper we take the kernel $\tilde{K}(u)=\frac{1}{2}|u|\exp(-|u|)$ where $u=(x_i-x_j)/h$. However, since these type of kernels are never positive (semi) definite, they cannot directly be applied with our smoother of choice i.e. an LS-SVM. Therefore, we review a two-step method developed by [1].

First, estimate the function with the NW estimator based on the kernel $\tilde{K}$ with bandwidth $\hat{h}_b$ (found by minimizing (5))

$$\hat{m}(x)=\sum_{i=1}^{n}\frac{\tilde{K}\left(\frac{x-x_i}{\hat{h}_b}\right)Y_i^{(1)}}{\sum_{j=1}^{n}\tilde{K}\left(\frac{x-x_j}{\hat{h}_b}\right)}. \tag{6}$$

From (6), we calculate the residuals

$$\hat{e}_i=Y_i^{(1)}-\hat{m}(x_i),\ \text{for } i=1,\ldots,n.$$

Now choose l to be the smallest $q\geq 1$ such that

$$|r_q|=\left|\frac{\sum_{i=1}^{n-q}\hat{e}_i\hat{e}_{i+q}}{\sum_{i=1}^{n}\hat{e}_i^2}\right|\leq\frac{\Phi^{-1}(1-\frac{\alpha}{2})}{\sqrt{n}}, \tag{7}$$

where Φ^{-1} denotes the quantile function of the standard normal distribution and α is the significance level, say 5%.

Second, once l is selected by (7), the tuning parameters of the LS-SVM (kernel bandwidth h and regularization parameter γ) can be determined by using leave-$(2l+1)$-out CV (see Definition 1) or modified CV combined with a positive definite kernel, e.g. Gaussian kernel.

Definition 1 (Leave-(2l + 1)-out CV) *Leave-$(2l+1)$-out CV or modified CV (MCV) is defined as*

$$\text{MCV}(h)=\frac{1}{n}\sum_{i=1}^{n}\left(Y_i^{(1)}-\hat{m}^{(-i)}(x_i)\right)^2, \tag{8}$$

where $\hat{m}^{(-i)}(x_i)$ is the leave-$(2l+1)$-out version of $m(x_i)$, i.e. the observations $(x_{i+j},Y_{i+j}^{(1)})$ for $-l\leq j\leq l$ are left out to estimate $\hat{m}(x_i)$.

To conclude this section, Algorithm 1 summarizes the model selection procedure for LS-SVM with dependent data.

Algorithm 1 Model selection procedure for LS-SVM with dependent data

1: Determine $\hat{h}_b$ in (6) with kernel $\tilde{K}$ by means of LOO-CV
2: Calculate l satisfying (7)
3: Determine both tuning parameters for LS-SVM by means of leave-$(2l+1)$-out CV (8) and a positive definite unimodal kernel.

4 Simulation

First, consider the following two functions $m(x) = 1 - 6x + 36x^2 - 53x^3 + 22x^5$ and $m(x) = \sin(2\pi x)$ with $n = 500$ equispaced points. The error variance was set to $\sigma^2 = 0.05$ and $e \sim \mathcal{N}(0,\sigma^2)$ for both functions. The value of k was tuned via LOO-CV and was set to 6 and 7 respectively for the first and second function. It is clearly shown that the proposed method is capable of estimating the first order derivatives nonparametrically quite accurate (see Figure 2).

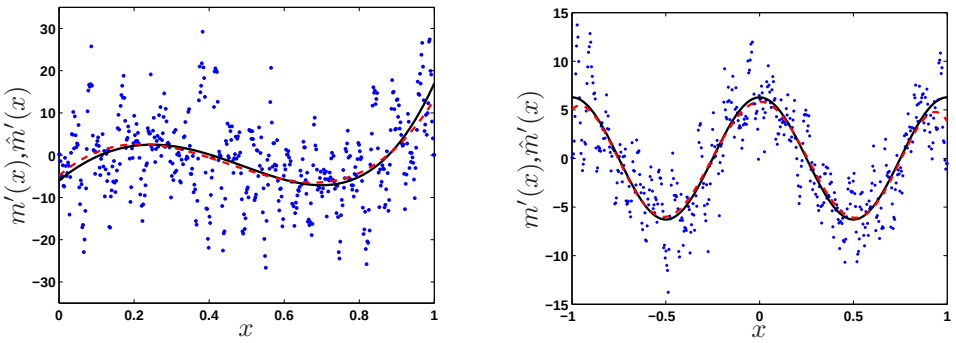


Figure 2: First order derivative estimation. Estimated derivative by the proposed method (full line) and true derivative (dashed line) for both functions. The value of k was tuned via LOO-CV and was set to 6 and 7 respectively for the first and second function.

Second, intuitively we could first estimate the regression function based on the given data set and then differentiate the LS-SVM regression estimate (3) w.r.t. to the independent variable. The next simulation shows that this idea does not always produce good estimates of the derivative function. Consider the function $m(x) = 1 + x\sin x^2$ (and hence $m'(x) = \sin x^2 + 2x^2\cos x^2$) with $n = 500$ equispaced points between $[-1,4]$. The error variance was set to $\sigma^2 = 0.05$ and $e \sim \mathcal{N}(0,\sigma^2)$. The value of $k = 5$ was found via LOO-CV. Further, we conduct the following Monte Carlo experiment. For 500 repetitions, we calculate the integrated absolute distance (IAD) between the true derivative m' and the two estimated versions of the derivative i.e., based on differentiating the estimated regression function $\hat{m}'_{\text{reg}}$ and the proposed derivative estimate $\hat{m}'$ respectively. Figure 3 shows a typical result of the estimates and Table 1 displays the average IAD and corresponding standard deviation for the experiment. This experiment clearly confirms the strength of the proposed method.

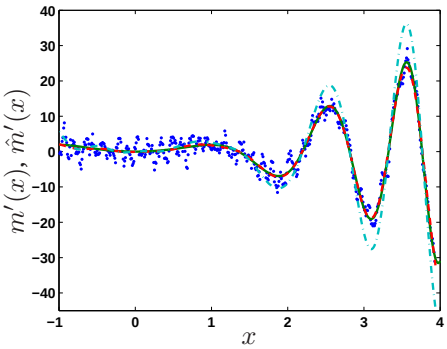


Figure 3: Comparison between the true derivative (full line), the derivative estimate based on the regression estimate (dash dotted line) and the proposed derivative estimate (dashed line).

IAD	$\int \lvert m'(x) - \hat{m}'_{\text{reg}} \rvert \, dx$	$\int \lvert m'(x) - \hat{m}'(x) \rvert \, dx$
average	13.36	1.78
standard deviation	0.32	0.047

Table 1: Integrated absolute distances and corresponding standard deviations for the experiment.

5 Conclusion

We proposed a simple but effective way of estimating derivatives nonparametrically via empirical first order derivatives. We have shown that this technique produces new data sets which are not independent and identically distributed (i.i.d.) random variables anymore. As an immediate consequence, all standard model selection criteria cannot be legitimately applied. We have illustrated how to modify the leave-one-out cross-validation so it is resistant against non i.i.d data. Finally, the method was illustrated on several toy examples.

Acknowledgements

Research supported by Research Council KUL: GOA/11/05 Ambiorics, GOA/10/09 MaNet, CoE EF/05/006 Optimization in Engineering (OPTEC) en PFV/10/002 (OPTEC), IOF-SCORES4CHEM, several PhD/post-doc & fellow grants; Flemish Government: FWO: FWO: PhD/postdoc grants, projects: G0226.06 (cooperative systems and optimization), G0321.06 (Tensors), G.0302.07 (SVM/Kernel), G.0320.08 (convex MPC), G.0558.08 (Robust MHE), G.0557.08 (Glycemia2), G.0588.09 (Brain-machine) research communities (WOG: ICCoS, ANMMM, MLDM); G.0377.09 (Mechatronics MPC), IWT: PhD Grants, Eureka-Flite+, SBO LeCoPro, SBO Climaqs, SBO POM, O&O-Dsquare, Belgian Federal Science Policy Office: IUAP P6/04 (DYSCO, Dynamical systems, control and optimization, 2007-2011), IBBT, EU: ERNSI; FP7-HD-MPC (INFSO-ICT-223854), COST intelliCIS, FP7-EMBOCON (ICT-248940), Contract Research: AMINAL, Other: Helmholtz, viCERP, ACCM. BDM is a full professor at the Katholieke Universiteit Leuven.

References

[1] K. De Brabanter, J. De Brabanter, J.A.K. Suykens, and B. De Moor. Kernel regression in the presence of correlated errors. *Journal of Machine Learning Research*, 12:1955–1976, June 2011.

[2] R. Charnigo and C. Srinivasan. Self-consistent estimation of mean response functions and their derivatives. *Canadian Journal of Statistics*, 39(2):280–299, 2011.

[3] P. Chaudhuri and J.S. Marron. SiZer for exploration of structures in curves. *Journal of the American Statistical Association*, 94(447):807–823, 1999.

[4] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2002.

[5] W. Härdle. *Applied Nonparametric Regression (reprinted)*. Cambridge University Press, 1999.

[6] A. Iserles. *A First Course in the Numerical Analysis of Differential Equations*. Cambridge University Press, 1996.

[7] C. Park and K.-H. Kang. SiZer analysis for the comparison of regression curves. *Computationsl Statistics & Data Analysis*, 52(8):3954–3970, 2008.

[8] J.O. Ramsay and B.W. Silverman. *Applied Functional Data Analysis*. Springer-Verlag, 2002.

[9] V. Rondonotti, J.S. Marron, and C. Park. SiZer for time series: A new approach to the analysis of trends. *Electronic Journal of Statistics*, 1:268–289, 2007.

[10] J.A.K Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle. *Least Squares Support Vector Machines*. World Scientific, Singapore, 2002.

[11] A.B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.